DOCUMENT RESUME

ED 082 763 LI 004 484

AUTHOR	Jewell, Sharon; Brandhorst, W. T.
TITLE	Search Strategy Tutorial; Searcher's Kit.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	10 Oct 73
NOTE	86p.; (3 references); ERIC Data Base Users Conference, Columbus, Ohio, October 10-12, 1973
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Computers; Data Bases; Data Processing; Educational Research; *Information Retrieval; Information Seeking; Relevance (Information Retrieval); *Search Strategies; Tutorial Programs
IDENTIFIERS	*Educational Resources Information Center; ERIC

ABSTRACT

The Educational Resources Information Center (ERIC) system's computerized data base was the focus of a three-hour tutorial session on search strategies. This document is the workshop manual used by the tutorial participants. The discussion of the input phase of a computer search covers identification of the user population, receiving the inquiry, and the types of services offered. The actual mechanics of searching includes general principles of good searching, search theory and general manipulative capabilities, and specific properties of the ERIC system that affect computer search capabilities. There is a practice session in which three searches are structured step-by-step. The output phase of a computer search includes a discussion of output formats, output evaluation, and statistical records-keeping. Eighteen technical notes discuss various aspects within each of these phases. Notes on the vocabulary improvement program for the "Thesaurus of ERIC Descriptors" are appended. (SJ)

FILMED FROM BEST AVAILABLE COPY

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 082763

SEARCH STRATEGY TUTORIAL

SEARCHER'S KIT

Prepared by

Sharon Jewell
Assistant Director
ERIC Clearinghouse on Library and Information Sciences

and

W.T. Brandhorst
Director
ERIC Processing and Reference Facility

for the

ERIC Data Base Users Conference
Columbus, Ohio
October 10-12, 1973

LI 004 484

SEARCH STRATEGY TUTORIAL

OUTLINE

I. INTRODUCTION

II. INPUT PHASE

A. User Population - Knowing your user group; User needs, expertise, background; Purpose of information; Application of information conveyed (teaching, research, student, parent, administrator); User education, relations; Mandate of information center/collection.

Technical Notes: None

B. Receiving the Inquiry/Question - Personal visit, telephone, letter/telegram; Search question negotiation (asking questions, completing forms, etc.); Determine area of interest by going from general to specific; Other parameters: volume of output desired, years covered, types of publications, recall vs. relevance, known "hits"; Use of subject specialists.

Technical Notes: #1. Search Negotiation. Understanding the Request

C. Types of Service Offered - Retrospective vs. Current Awareness; Manual search vs. Computer search; Referrals; Pre-prepared bibliographies; Telephone responses; Form responses; All-purpose packets of information.

Technical Notes: #2. Advantages of Computer Searching Over Manual Searching

#3. Two Modes of Searching: Retrospective Searching/ Current Awareness

III. MECHANICS OF SEARCHING

A. General Principles of Good Searching - Know your: data base, reference tools, search system (software); Make use of: user inputs, feedback, previous work, statistics.

Technical Notes: None

B. Search Theory and General Manipulative Capabilities - Basic logical operators; Boolean logic; Symbology; Venn diagrams; Truth Tables

Technical Notes: #4. Search Symbology (Operators, Venn Diagrams, etc.)

#5. Use of Parentheses (To Avoid Ambiguity)

#6. Weights. Sorting Output By Weight

#7. Arithmetic Operators

#8. Text/String Searching

#9. NOT Logic

C. Properties of the ERIC System - Data elements available for searching; Indexing vocabularies; Indexing practices; Descriptor/ Identifier frequency statistics; Reference tools.

Technical Notes: #10. Data Elements Available for Searching

#11. Levels of Generality and Specificity

#12. Major-Minor Index Terms

#13. Identifiers

#14. Importance of Knowing Descriptor Frequency (Posting) Statistics

#15. Common Descriptor Selection Problems (Examples)

IV. PRACTICE SESSION IN STRUCTURING SEARCHES

Search #1 - Text Editing

Search #2 - Social Studies Instruction

Search #3 - Criterion Referenced Tests

V. OUTPUT PHASE

A. Output Formats - Printout format (continuous or unitized); Data elements displayed (options); Callouts; Introductory explanatory matter; Administrative data (requester, date, search title; number of hits, search equation); Sequence (newest or oldest first; other sorts); CRT display.

Technical Notes: None

B. Output Evaluation - No "hits"; Relevance; Recall; Output volumes; Evaluation form; Other feedback techniques.

Technical Notes: #16. No Hits - What to Do

#17. Recall and Relevance

#18. Output Volumes. What is Too Little? What is Too Much? Hit Limits. Reverse Chronological Sort.

C. Statistics and Miscellaneous - Analysis of requests; Geographic distribution; Types of requester; Topical areas; Increasing throughput; Use of prior searches.

Technical Notes: None

APPENDIX A: Vocabulary Improvement Program

II. INPUT PHASE

SEARCH NEGOTIATION, UNDERSTANDING THE REQUEST

This topic may seem so obvious that nothing useful can be said about it. Nevertheless, it continues to be an underestimated factor in the conduct of a successful search service.

Too frequently the searcher is impatient to take the initial inquiry data available and "run with it". A superior search would have resulted if the searcher had first asked the user a series of basic questions. It is important to get from the user all the parameters he has to offer, e.g., alternative ways of describing the topic, closely related topics, does the topic have a geographic or institutional attachment, what years of publication are desired, what volume of output is desired, which is more important - recall or relevance, is he aware of any good documents on the subject already in the system, are any of the major authors who write on this topic known, are only certain academic or grade levels involved, etc.

Practice varies as to whether or not the reference center requires the user to state the inquiry in the standardized language of the system. Sometimes in order to save time and manpower, the user is asked, for example, to select terms representing the topic of interest from an authority list such as the ERIC Thesaurus. This is almost always dangerous in that the user is not fully familiar with the vocabulary, the definitions of terms, the ways that they have been used in indexing, etc. Forcing him to use the authority list restricts him and, in effect, lessens the flow of information from user to searcher. Unless it is essential for economic reasons to make the user perform some of the search labors, it is much more effective to ask him to state his inquiry in narrative form in his own language. Encourage an uncensored and unlimited description. This provides the searcher with the maximum raw material/clues/intelligence with which to help solve the problem posed by the inquiry.

If there is not voice contact between user and searcher, then obviously the inquiry is reduced to written form by the user before being submitted. This may or may not be true if there is face-to-face or telephonic contact. In the latter case the phrase "search negotiation" can be particularly apt. As the searcher asks the user to state the problem, what is then said can trigger questions by the searcher. As specifications are identified, the searcher can immediately react with the user, informing him as to whether the system can handle that aspect and, if not, what alternatives exist. For example, the user may specify "6 year olds" in his question. The searcher may inform him that the terms EARLY CHILDHOOD (covering 4-6) and CHILDHOOD (covering 7-12) are in use by the system and ask him which would be preferable in this case. The user may specify a disability, a grade level, and a curriculum area in his question. The searcher can determine which of these concepts is prime. If it is the disability, then the other factors should not be in a commanding and limiting position in the search. This kind of

immediate, real-time, negotiation can clearly lead to great refinement of the question. The things the user thought were so obvious they didn't need stating are elicited by the skillful questioning of the searcher. The improved understanding of the request usually leads to more accurate strategy and a user more satisfied with the end product.

ADVANTAGES OF COMPUTER SEARCHING OVER MANUAL SEARCHING

The purpose of this paper is to list some of the conditions that can alert a reference center to the possibility that a computer search to answer a particular query may be justified. Obviously there are many situations where a computer search is not justified and cannot compete in terms of time or cost with a simple straightforward manual approach. If someone is interested in what has entered the ERIC file on the subject of READING over the last quarter, the most efficient solution is the conventional one of going to the latest issues of RIE (perhaps using the last cumulative index), looking in the indexes, and perhaps photocopying a few pages.

There are other situations, however, where the computer can add a dimension to a search not obtainable manually.:

1. Multi-Factoral Searches

   This is a search involving more concepts (and therefore terms) than can reasonably be held in the mind, much less manipulated logically, while carrying out a manual search. For example, the intersection of three large "families" of terms, each involving perhaps 5-7 closely related terms, can result in a search requiring cognizance over a total of 20 terms or more. Specifically, imagine that the patron is looking for material on the use of innovative teaching tools (e.g., Audiovisuals) in non-public schools, particularly in the smaller schools, such as Church schools. The strategizing could easily result in the following kind of three family intersection:

PARAMETER A	AND	PARAMETER B	AND	PARAMETER C
INSTRUCTIONAL AIDS		INSTRUCTIONAL INNOVATION		CATHOLIC SCHOOLS
OR		OR		OR
INSTRUCTIONAL MATERIALS		INNOVATION		CATHOLIC ELEMENTARY SCHOOLS
OR		OR		OR
INSTRUCTIONAL MEDIA		EDUCATIONAL INNOVATION		PAROCHIAL SCHOOLS
OR				OR
INSTRUCTIONAL TELEVISION				PRIVATE SCHOOLS
OR				OR
TEACHING MACHINES				PROPRIETARY SCHOOLS

   The above search involves only 13 terms but one can readily see that to perform it manually would be impractical.

2. Large Files

   Any request where the size of the file to be searched is in the tens or hundreds of thousands is a potential candidate for the employment of the computer. The sheer clerical work involved in interrogating files of this size and recording the results argues for the use of that "super-clerk" the computer. The search may be a simple one or two term search. What one

is buying, therefore, is not logical or multi-factoral capability so much as the sheer convenience of letting the computer do the scanning, the selecting, and the assembling into a nice convenient package of the output.

3. Knowledge in New Patterns

It has been argued that computerized searching will more and more become a tool for those working with new configurations of knowledge. Because it provides the capability for doing highly complex searches, organized on any of the fields in a record, upon massive quantities of data, the machine search will promote the extraction of information in new patterns, rather than being merely a quicker way of doing old things. It is thought that the computer approach will be able to detect the coincidence of concepts that even the indexer may not have realized at the time; or that the computer will be able to detect statistical patterns across numerous accessions that would have escaped the human inputters of data simply because they work one item at a time. This argument has been made particularly strongly for those systems dealing with natural text, e.g., a system analyzing the complete text of the Dead Sea Scrolls; or for those systems dealing with large amounts of numerical data, e.g., Census Tapes analyzers.

4. Multiple Searches

Volume alone may argue for the computer approach. If a center must fulfill the search requests of many patrons during essentially the same period of time, the speed and accuracy of the computer can become powerful allies. Obviously heavy demands and "crash" demands are not new to service organizations, but as the pressures rise to extend services and increase productivity, without adding staff, the computer may provide a way out. This can extend both to the primary situation where the need is for single copies of different data (e.g., searches for 25 different professors preparing reading lists) and to the secondary situation where the need is for multiple copies of the same data (one search can be printed on multi-part paper or printed several times).

TWO MODES OF SEARCHING:

RETROSPECTIVE SEARCHING
CURRENT AWARENESS

Let us assume that a reference center has just acquired a new data base such as ERIC. The first item they received was the back-file of records as they existed at the time the order was processed. Later they receive smaller sections, update tapes (either monthly, quarterly, or annually) coming in at regular intervals.

Given this environment, we might project that the user would ask for one big retrospective search initially, followed by regular current searches of the update tapes, as they arrive.

1. Retrospective Searching

The retrospective search attempts to examine the entire data base comprehensively on a given topic. It is a large-scale effort usually done one time for any particular patron or topic. It requires careful search negotiation and coordination in order to be sure that it is precisely what the patron wants and that it will achieve the desired degree of completeness. The care is necessary because retrospective searches are typically fairly costly and have large outputs. It is prudent to process them carefully and to avoid wasteful mis-steps.

A typical problem in strategizing a retrospective search is to avoid excessive output. An ERIC search, for example, is being run against a file of about 150,000 accessions. Some of the more heavily posted terms have the ability to dump as many as 7,000 accessions in your lap, or 5% of the file. It is more likely that you are looking for no more than 150 hits (.1%). This means that your search must be tightly written and that it must take heed of the posting statistics for the terms that it is using. At the same time, as a comprehensive search, it must make certain that it utilizes all the terms that apply to the topic in question. Retrospective searches typically use a lot of terms and complicated logic, with intersections based on posting levels.

2. Current Awareness

In 1958, H. P. Luhn began to describe in the technical literature the pattern of service that came to be called first Selective Dissemination of Information (SDI) and which now tends to be referred to as "Current Awareness" searching. This involves, quite simply, the periodic running of a customized search for a particular individual against the latest data available. The search itself did not have to be re-submitted by the user; rather it was kept on file at the reference center. It was carefully tailored to fit the user's needs and might even have some extremely idiosyncratic characteristics, such as parameters relating to the user as author, the journals he subscribes to, the laboratory in

which he works, etc. This so-called "profile" of the user was regularly kept up-to-date by action of the center staff, the user himself, or both. Letting the user manage his own profile can be dangerous, but has the advantage of letting him "play the game", thereby involving him intimately in the information system and feedback to it.

The profiles are typically stored as a series of searches and run against incoming update tapes. Lancaster has made much of this by stating:

> "The principal distinction between SDI and retrospective searching systems is that in the case of the latter, a user request precipitates a search of the document file, whereas, in the former, a document precipitates a search of the user file".

The Current Awareness approach, using computers, increases the scale on which individually tailored services can be undertaken by a busy reference center. It also permits many refinements in service. Perhaps its most important contribution, however, has been that it represents an active dissemination of information, rather than a passive response. Librarians have often been critized for being mere preservers of records but Current Awareness fits in with the more dynamic and modern role of being specialists in the transmission of information to those who need it. Current Awareness takes the initiative rather than waiting for the user to come in the door.

A typical problem in strategizing a current awareness profile is to ensure that some output is achieved. If a profile is run monthly against the ERIC data base, it is searching only 1,000 - 1,300 records; if it is run quarterly, it is searching only around 3,500 records. Against such a small fraction of the entire data base it is necessary to structure a search rather loosely, in order to guarantee hits. Remember that even if the strategy were to dump 5% of the file (a disaster in retrospective searching), in Current Awareness searching, against a monthly tape, this would involve only 50 hits, an easily digestible quantity. Current Awareness profiles usually, therefore, involve a lot of OR logic and few AND statements. Posting data is relatively unimportant when constructing profiles. It is definitely not appropriate to simply take a retrospective search on the same topic and use it against update tapes without modification.

III. MECHANICS OF SEARCHING

III. MECHANICS OF SEARCHING

A. General Principles of Good Searching

1. Know the data base you are searching (what is in it; how it was built; etc.).

2. Know the search system capabilities available and how to use them most effectively.

3. Follow good search negotiation procedures with the requester, e.g.:

   a. Purpose for which information is to be used.

   b. Type of search - retrospective or current awareness.

   c. Amount of information expected - new or old, general or specific.

   d. Kind of information wanted - research, bibliographies, etc.

4. Use all reference tools available (including prior searches).

5. Make use of all search capabilities wherever possible.

6. Formulate strategy in terms of the user's request and expectations---avoid personal biases of the information retrieval specialist.

7. Evaluate output in terms of the original request.

8. Obtain feedback from the user in order to be able to improve service.

9. Keep statistics on user satisfaction, search results, etc., in order to improve service.

B. Search Theory, and General Manipulative Capabilities

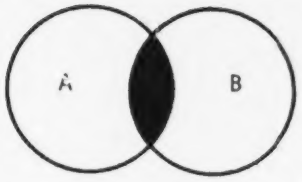
BOOLEAN LOGIC

BOOLEAN CONNECTOR/ OPERATOR	SYMBOL	ALGEBRAIC REPRESENTATION	MEANING
AND	. &	A . B A & B	Both A and B must be 'true' or must 'occur'.
OR	+ !	A + B A ! B	Either A or B, or both, must be 'true' or must 'occur'.
NOT	— ¬	A $\overline{B}$ A & ¬ B (A ¬ B)	A must be 'true' or must 'occur' and B must be 'not true' or must 'not occur'.

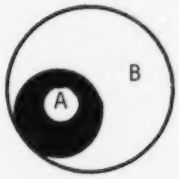
NOTE: In the above examples of symbols, the first version employs the traditional logical notation while the second shows conventional typgraphical symbols that can be used on keyboards to input the desired logic to the computer (e.g., via card-punches, video terminals, magnetic tape typewriters, etc.). Remember that + equals logical OR, not logical AND, and that it is an inclusive "OR not an exclusive "OR".

VENN DIAGRAMS

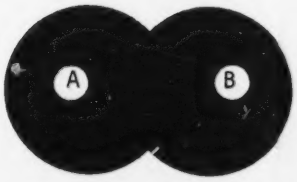
#1



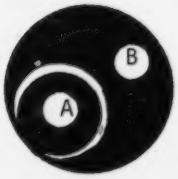
____ A & B ____



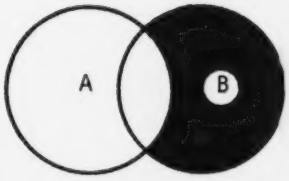
#2



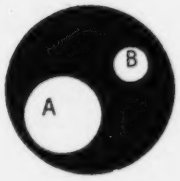
____ A | B ____



#3



____ B & ¬ A ____



ASSUME: A = Poems
B = Plays

#1 - Poems <u>and</u> Plays (only materials indexed with both terms)

#2 - Poems <u>or</u> Plays (all materials indexed with one or both terms)

#3 - Plays <u>but not</u> Poems (all materials indexed with the term Plays, excluding any indexed with the term Poems)

TRUTH TABLES

A	B	AND	OR	NOT
0	0	0	0	0
0	1	0	1	0
1	0	0	1	1
1	1	1	1	0

B: 0 | 1

A: 0 | 1

A \ B	0	1
0	AND = 0 OR = 0 NOT = 0	AND = 0 OR = 1 NOT = 0
1	AND = 0 OR = 1 NOT = 1	AND = 1 OR = 1 NOT = 0

1 = True or Present

0 = False or Not Present

SEARCH SYMBOLOGY (OPERATORS, VENN DIAGRAMS, ETC.)

There are a large number of symbols that have been used to represent the operations that search system designers wish to perform. There is little agreement on standard symbols even for the most common operators, the Boolean AND, OR, NOT operators. Because of this, many designers have preferred to use the words AND, OR, NOT rather than use symbols for them. Some use of arbitrary symbology is inevitable in any system, however, and the searcher must simply learn the language of the particular system he is involved with. In addition to the operators mentioned above, there may be symbols to indicate:

(1) that a root or string is being searched and not a whole word (See "Text/String Searching");

(2) that the word must appear as a major index term and not a minor (See "Major-Minor Index Terms");

(3) that two terms must appear adjacent to one another (See "Text/String Searching");

(4) that certain sub-files should be searched and not others;

(5) that any N of X terms listed are sufficient to generate a hit;

(6) that the output should be sorted in reverse chronological order (latest first);

(7) that only a set number of hits should be printed out;

(8) that the search should be saved in the system and be callable by instruction for future use, etc.;

(9) that the data found should be greater than or less than a certain preset value (see "Arithmetic Operators");

(10) that the "hits" should be sorted in order of potential relevance (See "Weighted Searching").

Figure A, attached, is an attempt to display, in an easy to reference manner, the various ways that the common Boolean operators can be represented.

Figure B, attached, is a comprehensive display of the representations of two terms in all possible logical combinations.

Both figures can be useful references for the active searcher who is not mathematically oriented and may occasionally have to verify what he is doing.

EQUIVALENT LOGICAL NOTATIONS

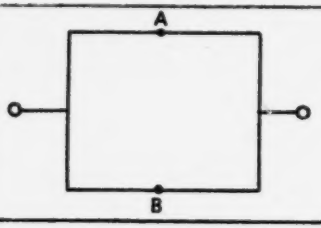
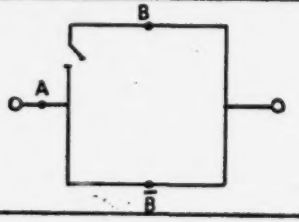
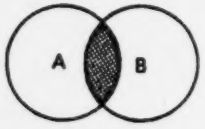
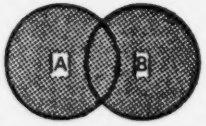
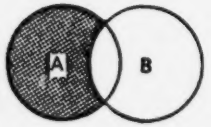
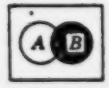
		AND	OR *	NOT **
WORDS		INTERSECTION MULTIPLICATION PRODUCT	COMBINATION SUM UNION	NEGATION
		CONJUNCTION	DISJUNCTION	COMPLEMENT
SYMBOLS		x e.g. A x B • A • B ∩ A ∩ B * A * B No Space AB Parentheses (A)(B)	+ e.g. A + B ∨ A ∨ B ∪ A ∪ B	- e.g. A - B $A\bar{B}$
GRAPHICAL REPRESENTATIONS	Switching Circuits			
	Venn Diagrams	 A and B	 A or B	 A not B
TABULAR REPRESENTATION (Table of possible values of a proposition, depending on values of the individual sets involved. 1 = True (or present) 0 = False (or not present)	Truth Table A B 0 0 0 1 1 0 1 1	0 0 0 1	0 1 1 1	0 0 1 0
NUMERIC (WEIGHTED/THRESHOLD) REPRESENTATION		A and B A = 1 B = 1 Threshold = 2	A or B A=1 B=1 Threshold=1	A not B A= 2 B=-1 Threshold = 2

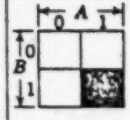
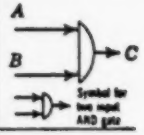
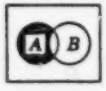
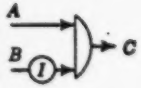
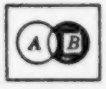
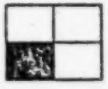
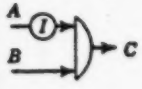
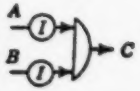
FIGURE A

* "OR" IS COMMONLY UNDERSTOOD AS THE "INCLUSIVE OR", I.E. A OR B OR BOTH. ALSO UTILIZED IS THE "EXCLUSIVE OR," I.E. A OR B BUT NOT BOTH, SOMETIMES SYMBOLIZED AS ⊕

** "NOT" IS EQUIVALENT TO "AND NOT" NOT "OR NOT"; IT IS SOMETIMES EXPRESSED "BUT NOT".

Logical Equation $C = f(A, B)$	Venn Diagram $C = f(A, B)$	Truth Table $C = f(A, B)$	Electronic Implemention $C = f(A, B)$
$C = A$			
$C = \overline{A}$			I = inverter
$C = B$			
$C = \overline{B}$			

(a)

Logical Equation $C = f(A, B)$	Venn Diagram $C = f(A, B)$	Truth Table $C = f(A, B)$	Electronic Implemention $C = f(A, B)$
$C = A \cdot B$			
$C = A \cdot \overline{B}$			
$C = \overline{A} \cdot B$			
$C = \overline{A} \cdot \overline{B}$			

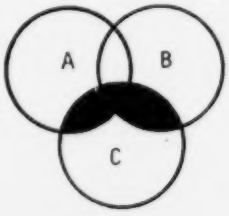
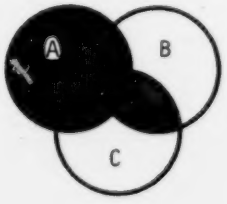
(b)

FIGURE B

USE OF PARENTHESES (TO AVOID AMBIGUITY)

Parentheses are a common way of indicating which terms in a search you want to have handled as one set. The ability to accept parentheses is a function of the software being used to access the file; it has nothing to do with the data itself. If a given search program does not accept parentheses, however, it either has to use a different symbol for the same purpose, or it has to have some conventions built into it to tell it in what order it is going to handle the terms and operators in the formulated query.

Parentheses remove the ambiguity that is otherwise present in a search equation. For example:

EQUATION WITHOUT PARENTHESES	POSSIBLE MEANINGS	GRAPHIC REPRESENTATION
A OR B AND C	1. (A OR B) AND C	1. 
	2. A OR (B AND C)	2. 

As can be seen, whether the search program performed the OR operation first or the AND operation first would make a great deal of difference as to what data were retrieved. The searcher can avoid any problems by telling the computer specifically in what order the terms should be combined. If the searcher leaves out parentheses (or their equivalent) the search program must either: (1) reject the query, stating that not enough information has been provided to interpret it properly, or (2) process the query according to previously agreed upon conventions; the usual conventions are that the program processes NOT, AND, OR, in that order.

PROBLEM

1. A OR B AND C OR D
2. A OR B AND C NOT D

These are ambiguous equations. Using the conventions referred to above, place parentheses around them to show how the computer would interpret them.

SOLUTION

1. A OR (B AND C) OR D
2. (A OR (B AND C)) AND NOT D

Note that if the search system did not permit parentheses, but used conventions instead, the searcher has no way of forcing the computer to treat equation #1 as (A OR B) AND (C OR D) and his searching is seriously restricted.

Remember that parentheses are meant to make explicit, not to confuse. Multiple sets of parentheses may look formidable, but they actually make things easier to figure out. Just begin on the inside, treating the contents of a set of parentheses like the contents of a small box that can be put inside another box. Always run a quick check by counting up the number of left parentheses and see that they are equal to the number of right parentheses. The counts must be equal for the equation to be logically correct.

EXAMPLE: ((A OR B) AND (C OR D)) AND (X OR Y)

Most search systems will reject a question in which the counts of left and right parentheses are not equal.

Parentheses give you power to specify exactly what you want done. A search system without parentheses (or their equivalent) would be highly limited indeed.

WEIGHTED SEARCHING

Searching may be accomplished by assigning weight values to index terms and then insisting that a document achieve a certain threshold weight value before it is considered a "hit". This approach is in use in several search systems both domestic and foreign. It has never seriously threatened the popularity of the basic Boolean approach, however, it does have certain advantages that have appealed to particular system designers. The two major advantages are as follows:

1. In a weighted search it is much easier to request that any N of X terms be present (e.g., any 2 of the 5 terms listed) to constitute a "hit"; this specification can be difficult and laborious to code using Boolean operators.

2. Weighted searching permits the searcher to arrange the output in order by weight value, thereby approximating an arrangement in order of relevance. In other words, weighting search terms injects a qualitative factor that the Boolean situation doesn't permit. Under Boolean operators, an item is either a "hit" or it isn't. Under weighted search conditions the resultant "hits" have varying weights and they can be arranged by these values.

Even though we are not aware of any search system employing weights that is currently accessing the ERIC data base, it is possible that this may change in the future. It may also be useful to ERIC searchers to understand the pros and cons of the weighted approach and how it relates to the Boolean approach.

The attached "Brief Communication" by a Facility staff member was originally prepared in 1966, however, we believe it still conveys the basic information that anyone contemplating the weighted search approach should know. For those who may wish to probe more deeply into the topic the most complete treatment yet prepared is: Pauline Angione's "On the Equivalence of Boolean and Weighted Searching Based on the Convertibility of Query Forms", M.A. Dissertation, Univ. of Chicago Graduate Library School, August 1968, 50 p.

Simulation of Boolean Logic Constraints Through the Use of Term Weights

The evolution described below of one aspect of the NASA Scientific and Technical Information Facility's machine search system may be of general interest to the documentation profession.

The Facility began operations in early 1962. The literature search service, or "demand bibliography" service, as it was then termed, was initially a very modest endeavor for the simple reason that the data base upon which to search had yet to be built. The first search programs concentrated on the well-known Boolean logic capabilities in the searching of inverted term files on magnetic tapes. This was consistent with the contractor's (Documentation Incorporated) prior R&D experience with so-called "Uniterms" and coordinate indexing systems.

A major change was effected, beginning in January 1965, to a serial or linear type of file organization. The reasons for this change were many and varied and need not concern us in any detail here. They involved, primarily, efficiencies in the file maintenance and update procedures and in the journal index preparation procedures. Also, it was becoming imperative to be able to search the file on a variety of nonsubject, administrative categories of information. At the time of this change, additional capabilities were built into the new "linear" search system. To supplement the basic Boolean capability, we now, among other things, made available to ourselves the following strategies that were well known in the state-of-the-art: (1) a weighting technique, (2) a "root" searching technique, and (3) a system of nonsubject "limits."

The weighting technique permits the assignment of arbitrary weight values to search terms and the specification of a minimum weight which any document must achieve in order to become a "hit."

"Root" searching permits queries on any desired generic level of various entities, e.g., all contracts with the prefix NAS8-; all report numbers with the prefix RAE-; all authors with names beginning CAR-. It may soon be extended to index terms, as in all terms beginning "PNEUMO," etc.

The system of "limits" permits the specification of various additional constraints on a search other than those involving subject index terms. Nearly all the standard descriptive cataloging elements fall within this system.

Each of these new capabilities has seen a great deal of use. The weighting technique, however, has particularly caught the interest of the searching staff and has resulted in some far-reaching developments.

For instance, it is apparent that document weight becomes a way of ranking search output in order of relevance. Probably the first use that weights were put to within the Facility was not to limit the output — the Boolean equation did this — but to *arrange* it for either the user or the analyst or perhaps both. This became extremely valuable in an environment where search output received a human edit before it was released. Arbitrary weight levels could be set by the analyst above which relevance to the question was assumed and below which his editorial effort was concentrated.

It also became apparent that the weighting technique could, by itself in some situations, achieve exactly the same results as a Boolean equation; cleverly assigned weights could *simulate* such an equation. For example, the equation (1) $A(B + C + D) =$ Answer, can be completely bypassed through the following weight assignments: $A = 3$, $B = 1$, $C = 1$, $D = 1$; Weight Limit $= 4$. This becomes very useful to know, for the calculation of weights was a much faster computer process than the solving of a Boolean equation, and the substitution could lead to significant computer time savings. Other common types of substitutions were the following:

(2) $A + B + C + D$	$A = 1, B = 1, C = 1, D = 1$ Weight Limit $= 1$
(3) $A \cdot B \cdot C \cdot D$	$A = 1, B = 1, C = 1, D = 1$ Weight Limit $= 4$
(4) $A + (B \cdot C \cdot D)$	$A = 3, B = 1, C = 1, D = 1$ Weight Limit $= 3$
(5) $(A + B) + (C \cdot D)$	$A = 2, B = 2, C = 1, D = 1$ Weight Limit $= 2$
(6) $(A + B) \cdot (C \cdot D)$	$A = 1, B = 1, C = 2, D = 2$ Weight Limit $= 5$

Various rules of thumb can easily be developed, and were, for the proper assignment of weights in more complex situations of the above basic types. However, no mathematical formalization was ever attempted.

It was soon realized that though term weighting had its advantages, nevertheless there were some equations that could not be reduced in this way. Two of the most basic are the following:

(7) $(A + B) \cdot (C + D)$

(8) $(A \cdot B) + (C \cdot D)$

The above equations cannot be simulated through any assignment to their terms of positive or negative weights, in conjunction with a weight limit. This can be proved by fairly simple algebraic techniques which will not be gone into here.

Continuing examination of the recalcitrant situations led to the development of a special "Group Weight" system for processing them. Essentially this involves "multiplying out" the equation, identifying its sections or groups, and assigning weights and weight limits for each section. Equation (7) thus becomes the redundant (7A) $A(C + D) + B(C + D)$ and weights may be assigned as follows:

Group A: $A(C + D)$	$A = 3, C = 1, D = 1$; Weight Limit $= 4$
Group B: $B(C + D)$	$B = 3, C = 1, D = 1$; Weight Limit $= 4$

The search program is now in the process of being changed to permit this technique. Logical equations will be made an optional, not a mandatory, feature of a search question. All types of logical equations may then be converted solely to a system of term weights and weight limits. Tests have been run comparing search times for ten problems coded by equation against the same ten coded with weights; both sets being run on our IBM-1410 search system against the same single reel of the data base. Results indicate that there is a 4 to 1 time advantage to running in the weight or arithmetic mode. However, it is clear that complicated equations can be both difficult and laborious to code. The next step is therefore obvious. In those cases where weights would be used mainly to simulate Boolean logic for the sake of processing speed, there is no reason that the program should not accept the equation and calculate its own weight assignments. This is now being evaluated.

It is thought that this particular case history in the use of weights may be of interest because of the widespread current use of weights in machine search systems. Several systems seem to be dropping the Boolean capability per se altogether in favor of weights. The two are generally spoken of in these situations as disparate entities. It is not that simple. The closeness of the relationship is shown by the fact that the weighting technique can be made to simulate Boolean logic. However, in doing so, the weighting technique can easily become too difficult for convenient human use. On the other hand, the logical equation is perhaps the most unambiguous and easily comprehensible way a search question with a complex relationship of terms can be organized and displayed. Our own solution is to keep both strategies in order to take advantage of the *unique* capabilities that each has to offer. At the same time, we are attempting to take advantage of the newly realized (at least as far as we are concerned) relationship between the two systems by utilizing the fast weight calculation process as a technique for internal computer solving of a logical equation.

W. T. Brandhorst
Assistant Director for Operations
NASA Scientific and Technical Information Facility
and
Documentation Incorporated
Bethesda, Maryland

ARITHMETIC OPERATORS

Arithmetic Operators are useful in the searching of data bases that have numerical fields or components, e.g., Census Tapes. These operators instruct the computer to proceed by making simple arithmetic tests of the data fields examined. Usualiy the test specifies that a certain range of data must be found in the field rather than specifying a specific number.

The common Arithmetic Operators are:

OPERATOR	SYMBOL	POSSIBLE ALPHABETIC REPRESENTATION
Equal To	=	EQ
Greater Than	>	GT
Less Than	<	LT
Greater Than or Equal To	$\geq$	GE
Less Than or Equal To	$\leq$	LE

Note that NOT logic in this environment can be handled by asking for the reverse condition. For example, if you wish to eliminate all hits with a publication date earlier than 1960 (NOT < 1960), this is equivalent to specifying that all hits have a publication date of Greater Than or Equal To 1960, i.e., $\geq$ 1960 is the reverse of < 1960.

Arithmetic Operators are not usually found in search systems designed to access bibliographic data. The reason is, of course, that the data elements found in such systems do not lend themselves to this type of handling. It must be stressed, however, that the availability of Arithmetic operators to a searcher is a function of the software at one's disposal; it is not a function of the data file.

TEXT/STRING SEARCHING

Most search systems, and certainly most search systems operating against the ERIC data base, rely on searching the Descriptor and Identifier fields for subject access to the file.

There are systems, however, which are designed to treat literally every word of the total record as a potential access point. These are generally called "full text" retrieval systems and they rely on "string searching" approaches. The "full text" usually refers to the full text of whatever is input as a record and not the full text of the actual document (which would be very expensive to key and store). What one finds, therefore, is that most "full text" systems are operating against the words in the title and abstract fields, as well as any indexing term field there may be. (Exceptions to this occur in the legal field, where the search may be against the actual full text of a statute).

Because such systems operate on natural unstandardized language, they are faced with all the problems caused by different endings and word forms, e.g., steal, steals, stealing, stealer, stolen, stole. This is why they generally provide string searching capabilities, permitting the searcher to specify given sequences of characters no matter where they appear, e.g., the root or string STEAL, no matter what ending it might have. They also take advantage of the fact that topics written about in close proximity to one another will generally be related. This is done by permitting the searcher to specify that A and B must appear in the same sentence for the item to be a hit; or they must appear within two words of each other, e.g., all items where INFORMATION and RETRIEVAL appear within two words of each other, as in "Information Storage and Retrieval". This is often called an "adjacency" capability.

On the theory that words mentioned early in an abstract are more important than words mentioned later on, some full text systems provide the capability of specifying that the terms must appear in the first so many sentences (or the first 50%) of the abstract.

The argument as to whether retrieval is better when relying on standardized index terms assigned by human indexers, or whether it is better when relying on the natural untouched text of the item itself, is sometimes called "The Great Debate" in information retrieval work.

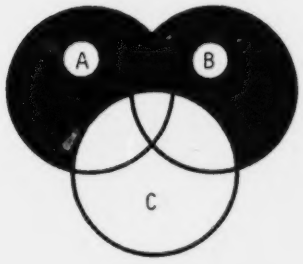
NOT LOGIC

Negation, or the exclusion of items because they have a particular property, is worth a short write-up because it is often either misunderstood or mis-applied. Some searchers are afraid to use it and never make it part of their armamentarium; others use it too much without realizing how much they might be missing as a result.

NOT logic usually has its own symbol and takes precedence in the hierarchy of machine operations. If a negative operator is interspersed in a logical equation with other operators, you can expect it to function first and most restrictively.

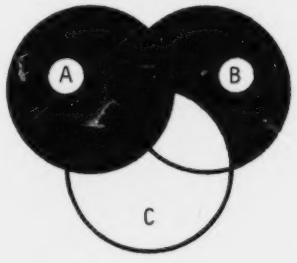
In other words:

Equation	Possible Interpretations	Normal Machine Interpretation
A OR B NOT C	(A OR B) NOT C A OR (B NOT C)	(A OR B) NOT C



(A OR B) NOT C

MOST RESTRICTIVE

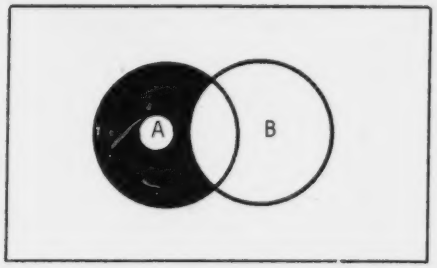


A OR (B NOT C)

LEAST RESTRICTIVE

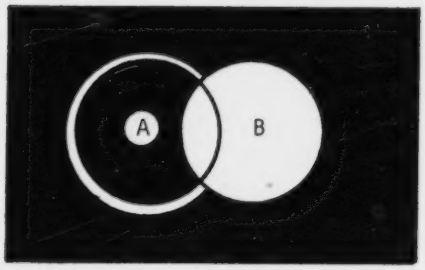
Even the most enthusiastic users of NOT logic admit that it can be a highly restrictive tool, often eliminating the good with the bad. Some recommend using it only after one search has already been done without it; using it then to eliminate known irrelevancies. Others recommend what is essentially the same thing, that the set of items being eliminated be examined to see just what is being lost. Both of these recommendations are recognitions of the fact that an item may meet every one of your positive specifications, but if it contains the single parameter negated, it can be excluded from the final printout.

It is of interest to note that NOT is short for AND NOT, not OR NOT:



A - B

A AND NOT B



A - B

A OR NOT B

DATA ELEMENTS AVAILABLE FOR SEARCHING

The machine-readable bibliographic records being manipulated, searched, and retrieved by various search systems, usually contain much more than just the subject index terms that tend to be concentrated on during searching. This is certainly true of ERIC and it holds true generally for the other major data bases as well.

The existence of these other data should not be forgotten. It is surprising how frequently they can be put to use with advantage in a subject search if the software being used permits it. In a search on BEHAVIOR MODIFICATION, for instance, surely it would pay to examine the works authored by B. F. Skinner. A search on ARTIFICIAL INTELLIGENCE might well add as a significant parameter the Institution (laboratory) by this name at the Massachusetts Institute of Technology. A search involving some aspect of JUNIOR COLLEGES might prefer to limit itself to input from the ERIC Clearinghouse on Junior Colleges in order to ensure high relevance for its subject search. A search on SPACE SCIENCES education might like to restrict itself to items having Report Numbers beginning NASA-EP, in order to pick up all the NASA educational publications in the ERIC system.

The above are just a few of the ways that non-subject data elements might come into play in what is basically a subject search. The ability to access these data depends completely on the software system being used. If the search system is a linear or sequential search against the ERIC Master Files, for instance, then it is necessary to pass all the data by the reading heads of the tape drives and the chances are that the system provides (or can be easily modified to provide) access to either or both non-subject and subject fields. If the search system first queries an inverted index file to determine the accession numbers of the "hits" that satisfy the specified conditions, then the first pass can involve only those data elements for which index files exist. This automatically excludes many of the non-subject data elements. However, in the second phase of such systems, it is necessary for them to go to the Master File and extract the full records for the "hits". It is sometimes possible to apply non-subject restrictions at this stage of the process, after there has already been a winnowing down on the basis of subject.

A complete list of ERIC data elements available for searching appears as Figure 1.

IX. ERIC DATA ELEMENTS AND THEIR FIELD IDENTIFICATION NUMBERS

A. RESUME SUBSYSTEM

ORDERED BY FIELD IDENTIFICATION NUMBER

DATA ELEMENT/FIELD	KEYWORD	USED IN*	HEXADECIMAL	DECIMAL	NOTES
Sequence	None	F J P R(ALL)	00	0	
Add Date (Julian)	None	F J P R(ALL)	01	1	
Change Date (Julian)	None	F P R	02	2	
*Accession Number	ACC	F J P R(ALL)	10	16	
*Clearinghouse Acc. No.	CH	J R	11	17	
Other Accession Number	OCH	R	12	18	
Publication Type	PUBTYPE	R	13	19	
Program Area	PA	P R	14	20	
Proposal Date	PROPDAT	P	15	21	
Project Date	PROJDAT	P	16	22	
Publication Date	PDAT	R	17	23	
Total Project Dollars	TOT	P	18	24	
Fiscal Year Funding	FYF	P	19	25	
Title	TITL	J P R	1A	26	
Personal Author	AUTH	J R	1B	27	
Institution	INST	F P R	1C	28	CODE #
Responsible Branch	BRANCH	P	1D	29	CODE #
Geographic Locality	GEO	F P	1E	30	CODE #
Cooperating Institution	CINST	P	1F	31	CODE #
Sponsoring Agency	SPON	P R	20	32	CODE #
Funding Agency	FUNDG	P	21	33	CODE #
Funded Individual	FUNDED	P	22	34	
Descriptors	DESC	F J P R(ALL)	23	35	
Identifiers	IDEN	F J P R(ALL)	24	36	
EDRS Price	PRICE	P R	25	37	
Descriptive Note	NOTE	R	26	38	
Project Officer	PRO	P	27	39	
Principal Investigator	PRI	P	28	40	
Field Reader	READ	F	29	41	
Address	ADDR	F	2A	42	
Issue	ISS	F J P R(ALL)	2B	43	
Abstract	ABST	F J P R(ALL)	2C	44	
Report Number	REPNO	R	2D	45	
Contract Number	CONT	F P R	2E	46	
Grant Number	GR	P R	2F	47	
Bureau Number	BN	P R	30	48	
Availability	AVAIL	J R	31	49	
Journal Citation	JNL	J R	32	50	
Experience Highlights	EH20-EH1	F	33-46	51-70	
Publications	PUBL	F	47	51	
Education	EDUC	F	48	72	
Region	REG	F	49	73	
Project Number	PN	P	4A	74	
Project Category	PCAT	P	4B	75	
Institution Name	None	F P R	80	128	28
Responsible Br. Name	None	P	81	129	29
Geographic Loc. Name	None	F P	82	130	(Derived 30
Cooperating Inst. Name	None	P	83	131	(From 31
Sponsoring Agency Name	None	P R	84	132	(Field 32
Funding Agency Name	None	P	85	133	33
Region Name	None	F	AD	173	73

ORDERED ALPHABETICALLY

DATA ELEMENT/FIELD	KEYWORD	USED IN*	HEXADECIMAL	DECIMAL	
Abstract	ABST	F J P R(ALL)	2C	44	
Accession Number	ACC	F J P R(ALL)	10	16	
Add Date	None	F J P R(ALL)	01	1	
Address	ADDR	F	2A	42	
Availability	AVAIL	J R	31	49	
Bureau Number	BN	P R	30	48	
Change Date	None	F P R	02	2	
Clearinghouse Acc. No.	CH	J R	11	17	
Contract Number	CONT	F P R	2E	46	
Cooperating Institution	CINST	P	1F	31	CODE #
Cooperating Inst. Name	None	P	83	131	NAME
Descriptive Note	NOTE	R	26	38	
Descriptors	DESC	F J P R(ALL)	23	35	
EDRS Price	PRICE	P R	25	37	
Education	EDUC	F	48	72	
Experience Highlights	EH20-EH1	F	33-46	51-70	
Field Reader	READ	F	29	41	
Fiscal Year Funding	FYF	P	19	25	
Funded Individual	FUNDED	P	22	34	
Funding Agency	FUNDG	P	21	33	CODE #
Funding Agency Name	None	P	85	133	NAME
Geographic Locality	GEO	F P	1E	30	CODE #
Geographic Locality Name	None	F P	82	130	NAME
Grant Number	GR	P R	2F	47	
Identifiers	IDEN	F J P R(ALL)	24	36	
Institution	INST	F P R	1C	28	CODE #
Institution Name	None	F P R	80	128	NAME
Issue	ISS	F J P R(ALL)	2B	43	
Journal Citation	JNL	J R	32	50	
Other Accession Number	OCH	R	12	18	
Personal Author	AUTH	J R	1B	27	
Principal Investigator	PRI	P	28	40	
Program Area	PA	P R	14	20	
Project Category	PCAT	P	4B	75	OBSOLETE
Project Date	PROJDAT	P	16	22	
Project Number	PN	P	4A	74	
Project Officer	PRO	P	27	39	
Proposal Date	PROPDAT	P	15	21	
Publication Date	PDAT	R	17	23	
Publication Type	PUBTYPE	R	13	19	
Publications	PUBL	F	47	71	
Region	REG	F	49	73	CODE #
Region Name	None	F	AD	173	NAME
Report Number	REPNO	R	2D	45	
Responsible Branch	Branch	P	1D	29	CODE #
Responsible Br. Name	None	P	81	129	NAME
Sequence	None	F J P R(ALL)	00	0	
Sponsoring Agency	SPON	P R	20	32	
Sponsoring Agency Name	None	P R	84	132	
Title	TITL	J P R	1A	26	
Total Project Dollars	TOT	P	18	24	

F = Personnel; J = Journal Articles; P = Projects, R = Reports

LEVELS OF GENERALITY AND SPECIFICITY

It is a generally established practice, in systems (like ERIC) employing coordinate indexing principles, that documents should be indexed at the level of specificity of the document in hand. In other words, the most accurate term in the Thesaurus that represents the concept covered by the document should be selected, not a term higher or lower in the hierarchy. For example, if a document deals with HANDICAPPED CHILDREN then that term should be selected rather than the broader term CHILDREN (perhaps coordinated with HANDICAPPED). However, if the document referes to all kinds of children, and handicapped children do not stand out as a distinct topic, then the broad term CHILDREN would be most appropriate. If the document treats both children generally and handicapped children specifically, then both CHILDREN and HANDICAPPED CHILDREN are appropriate Descriptors, even though both are in the same generic tree.

There are some systems which practice "automatic posting up". In other words, if a document is indexed at some middle point in a generic tree, such as INTELLIGENCE TESTS, it will also, as a matter of course, be indexed by its Broader Term TESTS. When this practice is followed, unless the indexing is tagged in some way, there is no way to distinguish between general materials on TESTS and more specific materials on INTELLIGENCE TESTS which have also been indexed to TESTS. The practice of posting solely to the levels actually dealt with by the document has the advantage of permitting the searcher to zero in with greater accuracy on the topic desired by the user. Conversely, however, it means that if the searcher is interested in retrieving at all levels of a given topic, it is necessary to include not only the broad generic term covering the area, but also the many specific terms lower in the tree. This can sometimes present the searcher with an onerous coding task. For example, under the term AFRICAN LANGUAGES in the ERIC Thesaurus, there are over 30 specific languages, such as SWAHILI. If a searcher is interested in everything the system has on African languages, whether general or specific, he must code all 30 terms into the search. Sometimes sophisticated search systems avoid this problem by permitting the searcher to specify a given term, plus all terms narrower to it. In other words, with one search instruction the searcher could pick up all the specific African languages without having to write each one down.

It is important to the searcher to be aware of indexing practice in this area. Let us assume, for example, the following hypothetical indexing situations:

1. Document discusses a general concept (e.g., SALARIES) but illustrates profusely from a narrower class (e.g., TEACHER SALARIES). Both are selected by the indexer.

2. Document concentrates on a specific concept (e.g., PSYCHOLINGUISTICS), but the indexer thinks the treatment is such that it adds useful information to the body of knowledge about the more general concept (e.g., LINGUISTICS); both terms are used.

3. Document discusses many specific concepts (e.g., NURSES, PHYSICAL THERAPISTS, DENTAL HYGIENISTS, etc.), but none in sufficient detail to merit the indexing of each specific concept. Instead, the generic term HEALTH OCCUPATIONS is used.

4. Document provides detailed treatment of several types of Agricultural Personnel (e.g., EXTENSION AGENTS, AGRICULTURAL LABORERS, FARM MECHANICS, SHARECROPPERS, FORESTRY AIDES, AGRICULTURAL TECHNICIANS, etc.). In the judgment of the indexer there is sufficient data on each to warrant indexing each specific occupational group. In addition, because there are so many groups involved, the general AGRICULTURAL PERSONNEL is used. (If only two or three types had been treated, the generic term would probably not have been appropriate).

5. Document deals solely with a specific test called the "Detroit Advanced Intelligence Test". The indexer thinks the document should be made accessible via Descriptor (as well as the specific Identifier) and chooses the "reasonable" level INTELLIGENCE TESTS (not TESTS).

6. Document is a comprehensive treatment of SUICIDE among all classes of people, including STUDENTS. The slant is specifically SUICIDE and therefore the Broader Term DEATH is not used.

All of the above solutions are justified under the ERIC guideline to index to the specific topic dealt with by the document. As can be seen, the indexer is given great discretion to interpret the subleties and emphases of the document. The searcher must be aware of the possibilities both in order to search effectively and to interpret search results.

MAJOR-MINOR INDEX TERMS

At the time it is used to index a document or article in the ERIC system, every Descriptor or Identifier is identified as representing either a "Major" concept in the document or a "Minor" concept. Vocabulary terms are not, therefore, major or minor in themselves, but only as they are applied in a given situation. For example, a document dealing basically with NURSERY SCHOOLS may touch peripherally on TOYS, as one factor to consider. *NURSERY SCHOOLS is, therefore, considered the "Major" concept and is identified as such by being tagged with an asterisk, as shown. TOYS, given much lighter treatment, is considered a "Minor" concept and is identified as such by the absence of an asterisk.

The following table shows the average number of Descriptors and Identifiers assigned to each RIE and CIJE accession and the proportion of these that are identified as Major and Minor:

	RIE	CIJE
Average Total Number of Descriptors Assigned to Each Accession	11.35	6.61
Major Descriptors	4.91	3.88
Minor Descriptors	6.44	2.73
Average Total Number of Identifiers Assigned to Each Accession	.97	.39
Major Identifiers	.18	.17
Minor Identifiers	.79	.22
Average Total Number of Index Terms (Both Descriptors and Identifiers) Assigned to Each Accession	12.32	7.00
Major Term	5.09	4.05
Minor Term	7.23	2.95

This practice of distinguishing between Major and Minor index terms serves two principal functions:

1. It Limits the Size of the Published Subject Index

In order to provide indexing in depth of all concepts covered significantly by an accession, an average of 12.32 total terms are assigned to each RIE accession and an average of 7.00 total terms are assigned to each CIJE accession. At the present time only the Major terms are permitted to appear in the published Subject Indexes. If all of the terms were permitted to appear, these indexes would be over twice their present size. Can you imagine an RIE Annual Index twice its present size? This would be impractical from a publishing and economic standpoint. The Major-Minor dichotomy permits the ERIC system to have the benefits of both in-depth indexing together with practical, reasonably large, published subject indexes.

2. It Permits Searchers To Go After High Recall or High Relevance (Precision)

If a searcher is interested in comprehensiveness, in getting everything in the system that touches on a subject, he can search on all the appearances of a term, without regard for Major or Minor. On the other hand, if the searcher wants only material that devotes itself heavily to the topic in question, he can restrict the search to the asterisked appearances of the term involved.

If the indexers had not made the Major-Minor distinction at input time, all the index terms would be on the same footing and the searcher would not be able to tell the key subjects from the peripheral subjects.

IDENTIFIERS

There are two types of indexing terms used in the ERIC system: Descriptors and Identifiers. Descriptors are tightly controlled, defined, and cross-referenced, and appear in the Thesaurus of ERIC Descriptors. They represent relatively well-known subject matter concepts such as ANTHROPOLOGY, NURSERY SCHOOLS, TEACHING MACHINES, etc. Identifiers represent virtually anything else that an indexer might like to subject index a document by. The Identifier field is meant to be a very open and unconstrained field giving the indexer great freedom and nearly complete discretion to include index access points that are deemed useful to the user.

Identifiers are, in almost all cases, the names of specific entities. As there is a nearly infinite number of specific entities, it is not appropriate to burden a thesaurus with such a multiplicity of entries. Also, Identifiers, being so specific and often transitory, may be represented in the literature very infrequently; this fact also argues for separate treatment.

The major purpose of Identifiers is to provide additional indexing depth, of a specialized nature, supplementing that provided by Descriptors. Identifiers may be specific projects, geographic locations, persons, trade names, tests, legislation, organizations, equipment, etc. It is also possible to use the Identifier category as a testing ground for a term whose permanence may be in some doubt. If the term demonstrates over time its acceptability by the profession, it may graduate from Identifier to Descriptor status, e.g., Computer Assisted Instruction (CAI). Identifiers are not defined (scoped), cross-referenced, structured (related to one another), or otherwise subjected to lexicographic analysis. In order to aid retrieval, however, it is necessary to observe certain standards in their construction and to see that the more frequently used ones appear in the file in a uniform format. The ERIC Processing Manual includes, as Appendix G, a list of the more heavily used Identifiers, in their preferred format.

The following is a list of the major categories of Identifiers, together with an example of each:

CATEGORY	EXAMPLE
Acronyms	PERT
Coined Terminology	Sesame Street
Conferences/Meetings/Seminars/Symposia	National Reading Conference
Equipment	Autotutor
Ethnic Groups/Tribes	Shoshones
Geographic Locations	New York City
Legislation	Taft Hartley Act
Methods and Theories	Montessori Method

Organizations

Community Organizations.....Los Angeles Chamber of Commerce
Educational Organizations.....Parent Teacher Association
Foundations.....Ford Foundation
Government Agencies.....National Institute of Education
Industrial Organizations.....Westinghouse Corporation
School Districts.....Milford Kansas School District

Personal Names.....Skinner (B F)
Projects.....Project Talent
Tests and Testing Programs.....Scholastic Aptitude Test
Textbooks.....Uralic and Altaic Series
Trade Names.....Erictapes

Note that Identifiers, like Descriptors, are tagged at indexing time as representing Major or Minor concepts in the document being processed.

The following notice, which appeared in Interchange #3 illustrates well how Identifiers can play a large role in a search, supplementing the Descriptors.

BRITISH INFANT SCHOOL - SEARCH STRATEGY

Carolyn Trohoski of RISE writes that a search for material in the ERIC system on the subject of the British Infant Schools requires the use of numerous Identifiers as well as Descriptors. The terms she used in her search, and that she finds are worth passing on to others, are shown in the table below. If it is desired to limit output solely to actual British references to these schools, as opposed to U. S. applications of the same principals, the searcher should intersect with the geographic Identifiers: ENGLAND or GREAT BRITAIN or UNITED KINGDOM.

TERM	DESC-RIPTOR	IDEN-TIFIER	RIE	CIJE
1. British Infant School		x	x	x
2. British Infant School Theory		x		x
3. British Infant Schools		x	x	x
4. British Primary Schools		x		x
5. Informal British Infant Schools		x	x	
6. Informal British Schools		x		x
7. Infant Schools		x	x	x
8. Leicestershire Infant Schools		x	x	
9. Open Classrooms		x		x
10. Open Education	x	x	x	x
11. Open Education Model		x		x
12. Open Education System		x	x	
13. Open Plan Schools	x		x	x
14. Open School		x	x	
15. Open Schools		x	x	x

IMPORTANCE OF KNOWING DESCRIPTOR FREQUENCY (POSTING) STATISTICS

In most information storage and retrieval systems the subject index terms display an enormous variation in the frequency with which they are used. In the ERIC system, for example, there is one term (INSTRUCTIONAL MATERIALS) that has been used over 4,000 times. There are 4 terms that have been used over 3,000 times. On the other hand, there are 136 terms that have been used only once. The attached Figure A gives some indication of the spread of the terms over the various usage levels.

It is absolutely essential for a searcher to know the usage levels for the terms being used in a search. It is possible to mismatch Descriptors so that the possibility of there being any hits becomes very poor. For example, Figure B depicts a situation where there is a total file amounting to 100,000 references. Term A has been used 1,000 times, Term B has been used 500 times; Term C has been used 5 times. If the assumption is made that the usages of these terms are equally likely to be scattered across any item in the file then the chances of there being an item containing both A and B is the multiplication of their separate probabilities, i.e.,

Chance of A appearing is $\frac{1,000}{100,000} = \frac{1}{100}$

Chance of B appearing is $\frac{500}{100,000} = \frac{1}{200}$

Chance of <u>both</u> A and B appearing is $\frac{1}{100} \times \frac{1}{200} = \frac{1}{20,000}$

With a probability of $\frac{1}{20,000}$ and a file size of 100,000, the anticipated number of hits would be $\frac{1}{20,000} \times \frac{100,000}{1} = 5.$

However, to intersect A and B and C would be to decrease the probability of any hits to essentially zero. For example:

Chance of A = $\frac{1}{100}$

Chance of B = $\frac{1}{200}$

Chance of C = $\frac{5}{100,000} = \frac{1}{20,000}$

Chance of A and B and C = $\frac{1}{100} \times \frac{1}{200} \times \frac{1}{20,000} = \frac{1}{400,000,000}$

* As of June 1972

Because of the mismatch between the usage level of C and the other two terms, the chances of a hit involving A and B and C are one in 400 million, or essentially no chance at all. C should not be intersected with any other term; its postings should probably be examined in their entirety.

Any examination into the probabilities involved in information retrieval makes one realize rather quickly that the very broad or general terms, with a high frequency of postings, that many people are fond of saying are of no use for retrieval, do indeed have their value in the context of machine searching. If indexers over a period of years have used INSTRUCTIONAL MATERIALS over 4,000 times, it is plan that this is one of the central topics appearing in the ERIC literature. By intersecting such a term with other, less frequent terms, the term can definitely serve as a filter and its high volume of postings is not necessarily a liability when the comparisons are being done by a high speed computer rather than a human being.

ERIC DESCRIPTOR USAGE

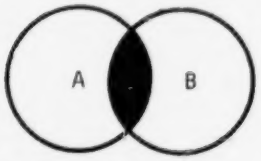
(Distribution of Postings by Various Ranges)

Range of Postings	1970 (June)		(1971 (June)		1972 (June)	
	Number of Terms	Percentage	Number of Terms	Percentage	Number of Terms	Percentage
1	229	5.08	184	3.87	136	2.81
2 - 3	337	7.48	301	6.33	250	5.16
4 - 9	626	13.89	570	11.98	522	10.77
10 - 49	2,403	53.30	1632	34.30	1608	33.18
10 - 99			2410	50.65	2392	49.36
50 - 99			778	16.35	784	16.18
100 - 199	534	11.85	651	13.68	725	14.96
200 - 299	205	4.55	272	5.72	319	6.58
300 - 399	74	1.64	143	3.01	173	3.57
400 - 499	32	0.71	70	1.47	106	2.19
500 - 599	23	0.51	58	1.22	57	1.18
600 - 699	10	0.22	20	0.42	52	1.07
700 - 799	7	0.16	20	0.42	24	0.49
800 - 899	6	0.13	13	0.27	20	0.41
800 - 999	5	0.11	3	0.06	10	0.21
1,000 - 1,999	15	0.33	33	0.69	45	0.93
2,000 - 2,999	2	0.04	9	0.19	10	0.21
3,000 - 3,999	0	0	1	0.02	4	0.08
4,000 +	0	0	0	0	1	0.02
Totals	4,508	100.00	4,758	100.00	4,846	100.00

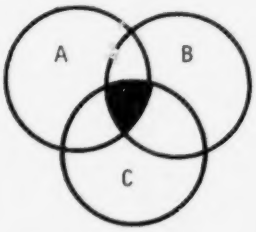
FIGURE A

Probability of each term appearing in record:

$\frac{1,000}{100,000} = \frac{1}{100}$ $\frac{500}{100,000} = \frac{1}{200}$ $\frac{5}{100,000} = \frac{1}{20,000}$



Documents indexed by A and B can be expected to be about $\frac{1}{100} \times \frac{1}{200} = \frac{1}{20,000}$ or, for a file of 100,000 references, about 5 records.



Documents indexed by A and B and C can be expected to be about $\frac{1}{100} \times \frac{1}{200} \times \frac{1}{20,000} = \frac{1}{400,000,000}$ or, for a file of 100,000 references, about .0002 records.

NOTE: Even if A and B were very likely to appear together and the number of records having both of them were 100 instead of 5, the probability of A and B and C all appearing together is still remote:

PROBABILITY OF A AND B: $\frac{100}{100,000} = \frac{1}{1,000}$

PROBABILITY OF A AND B AND C: $\frac{1}{1,000} \times \frac{1}{20,000} = \frac{1}{20,000,000}$

or, for a file of 100,000 references, about .005 records.

FIGURE B

COMMON DESCRIPTOR SELECTION PROBLEMS (EXAMPLES)

1. No Descriptor Representing Concept

DIFFERENTIAL PSYCHOLOGY	PARAPSYCHOLOGY
INSTINCTS	SLEEP TEACHING
LEARNING CENTERS	SPORTS (SPECIFIC SPORTS)

2. Descriptor Best Found Via Other Displays (e.g., Rotated)

EIGHTEENTH CENTURY LITERATURE

TWENTIETH CENTURY LITERATURE

NOTE: No entries in alphabetic display under CENTURY

VISUALLY HANDICAPPED

NOTE: Cannot be found under HANDICAPPED in alphabetic display because treated as NT to PERCEPTUALLY HANDICAPPED

LOCAL HOUSING AUTHORITIES

MALAYO POLYNESIAN LANGUAGES

STUDENT SCIENCE INTERESTS

3. Low Posted Descriptors (Must Not Intersect)

HORSES (2)
METALLURGICAL TECHNICIANS (1)
PARKING METERS (1)
SEISMOLOGY (2)

4. Descriptor Too Specialized (?)

ANISEIKONIA
AUDITORY AGNOSIA
CORN
HAGIOGRAPHIES
HETEROPHORIA
HIGH INTEREST LOW VOCABULARY BOOKS
HORIZONTAL TEXTS
ONOMASTICS
TAGMEMIC ANALYSIS
TRANSFORMATION GENERATIVE GRAMMAR

6. Descriptor Not Defined

ART SONG	INNER SPEECH (SUBVOCAL)
ARTICULATION (PROGRAM)	INPUT OUTPUT ANALYSIS
CONCEPTUAL SCHEMES	MILIEU THERAPY
CONNECTED DISCOURSE	NOMINALS
COORDINATION COMPOUNDS	NON GRADED CLASSES
DEEP STRUCTURE	UNGRADED CLASSES
DISTINCTIVE FEATURES	SERVICE OCCUPATIONS
INDIVIDUAL PSYCHOLOGY	

7. Descriptor Very Broad (Must be Intersected)

ABILITY	METHODS
ATTITUDES	NEEDS
BACKGROUND	OBJECTIVES
BEHAVIOR	PERFORMANCE
DATA	PLANNING
DEVELOPMENT	PROBLEMS
EDUCATION	PROGRAMS
ENVIRONMENT	RESEARCH
EVALUATION	SCIENCES
GROUPS	STUDY
GUIDES	TEACHING
INSTRUCTION	TECHNIQUES
LEARNING	THEORIES

8. Descriptors So Close in Meaning That They Must Be Used Together (Near Synonyms)

{HEREDITY
GENETICS

{NONFARM YOUTH
YOUTH

{EDUCATIONAL TELEVISION
INSTRUCTIONAL TELEVISION
TELEVISED INSTRUCTION

{INFORMATION RETRIEVAL
INFORMATION SEEKING
INFORMATION NEEDS
INFORMATION PROCESSING
INFORMATION SERVICES
INFORMATION DISSEMINATION
etc.

{COUNSELING CENTERS
GUIDANCE CENTERS
etc.

{EDUCATIONAL COUNSELING
EDUCATIONAL GUIDANCE
etc.

{EDUCATIONAL NEEDS
EDUCATIONAL OBJECTIVES

{EVALUATION METHODS
EVALUATION TECHNIQUES

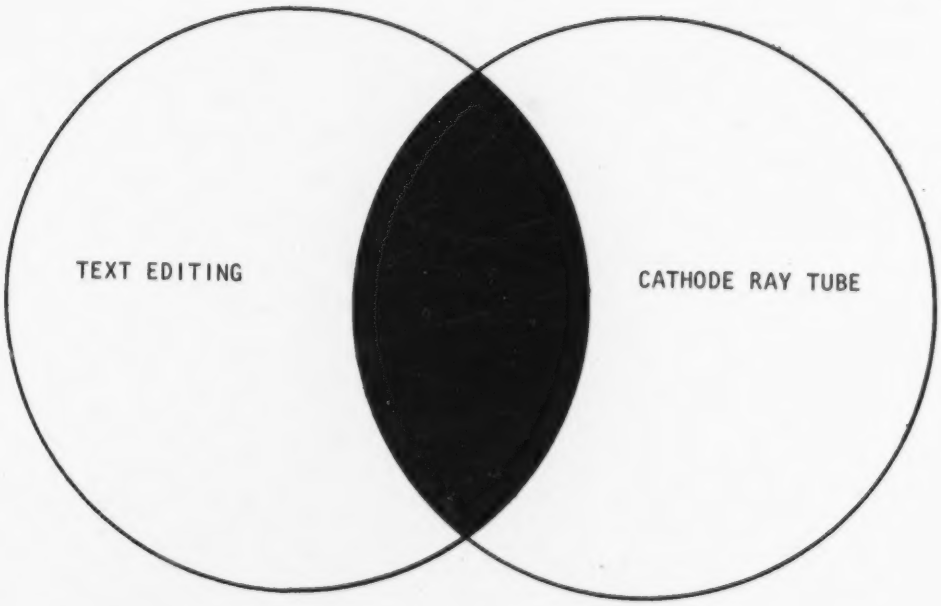
{PRE-SCHOOL PROGRAMS
PRE-SCHOOL EDUCATION
PRE-SCHOOL CURRICULUM

IV. PRACTICE SESSION IN STRUCTURING SEARCHES

SEARCH #1

TEXT EDITING BY CATHODE RAY TUBE

This is a simple, two-phase or two-family search combining the concepts "Text Editing" AND "Cathode Ray Tube".



Let's take the "Cathode Ray Tube" family of terms first.

CATHODE RAY TUBE
FAMILY OF TERMS

ERIC Thesaurus

ERIC Thesaurus Entries for Terms Suggested By RIE Indexing

CATHARSIS 060
SN RELAXATION OF EMOTIONAL TENSION BY EXPRESSIVE REACTION
UF ABREACTION
PSYCHOCATHARSIS
BT EMOTIONAL EXPERIENCE
RT AGGRESSION
ANXIETY
EMOTIONAL DEVELOPMENT
HOSTILITY
PSYCHOLOGICAL PATTERNS
PSYCHOTHERAPY
REACTIVE BEHAVIOR
SELF EXPRESSION

← No Descriptor

CATHOLIC EDUCATORS 380
BT TEACHERS
RT CATHOLICS
CATHOLIC SCHOOLS
CHURCH RELATED COLLEGES
CHURCH WORKERS
NUN TEACHERS
RELIGIOUS EDUCATION

COMPUTER ORIENTED PROGRAMS 270
SN THE APPLICATION OF COMPUTER TECHNOLOGY TO EDUCATION FOR BOTH INSTRUCTIONAL AND BUSINESS APPLICATION
BT COMPUTER PROGRAMS
RT COMPUTER ASSISTED INSTRUCTION
COMPUTERS
COMPUTER SCIENCE
COMPUTER SCIENCE EDUCATION
EDUCATIONAL TECHNOLOGY
ELECTRONIC DATA PROCESSING ←
PROGRAMED INSTRUCTION
TIME SHARING

DISPLAY SYSTEMS 050
BT INFORMATION SYSTEMS
RT COMPUTERS
ELECTRONIC DATA PROCESSING
ELECTRONIC EQUIPMENT
INFORMATION PROCESSING
INPUT OUTPUT
MAN MACHINE SYSTEMS ←
SCREENS (DISPLAYS)

ELECTRONIC DATA PROCESSING 080
SN DATA PROCESSING BY MEANS OF COMPUTERS
UF ADP
AUTOMATIC DATA PROCESSING
EDP
BT COMPUTER SCIENCE
DATA PROCESSING
RT AUTOMATION
COMPUTER ORIENTED PROGRAMS
COMPUTERS
COMPUTER SCIENCE EDUCATION
COMPUTER STORAGE DEVICES
DATA BASES
DATA PROCESSING OCCUPATIONS
DISPLAY SYSTEMS
INFORMATION SYSTEMS
INPUT OUTPUT DEVICES
ON LINE SYSTEMS
OPTICAL SCANNERS
PROGRAMING
PROGRAMING LANGUAGES

INPUT OUTPUT DEVICES 170
UF INPUT DEVICES
OUTPUT DEVICES
NT OPTICAL SCANNERS
BT EQUIPMENT
RT COMPUTER OUTPUT MICROFILM
COMPUTERS
COMPUTER STORAGE DEVICES
ELECTRONIC DATA PROCESSING
ELECTRONIC EQUIPMENT
FACSIMILE TRANSMISSION
INFORMATION PROCESSING
INPUT OUTPUT
MAGNETIC TAPES
ON LINE SYSTEMS
TELECOMMUNICATION

MAN MACHINE SYSTEMS 080
SN MEN AND MACHINES INTERACTING TO FORM SINGLE SYSTEMS
UF MAN MACHINE COMMUNICATION
MAN MACHINE INTERACTION
MAN MACHINE INTERFACE
RT ADVANCED SYSTEMS
AUTOMATION
BIONICS
COMPUTER ASSISTED INSTRUCTION
CYBERNETICS
DIAL ACCESS INFORMATION SYSTEMS
DISPLAY SYSTEMS
FEEDBACK
HUMAN ENGINEERING
INTERACTION
MANAGEMENT SYSTEMS
ON LINE SYSTEMS

Identifier Usage Report

Catell Intant Scale	ED026109
Cathode Ray Tube	ED047504 ED047524 EP012002
CATHOLIC CHURCH	ED019156
CRS	ED022636
CRT	ED047504 ED047524
Crusher Inspector	ED061302

An Accession Indexed by CATHODE RAY TUBE

ED 047 504 EM 008 705

Thomas, David B.

Two Applications of Simulation in the Educational Environment. Tech Memo.

Florida State Univ., Tallahassee. Computer-Assisted Instruction Center.

Spons Agency—Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.

Report No—AD-718-847; TM-31

Pub Date 71

Note—27p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, N.Y., February 4-7, 1971)

Available from—National Technical Information Service, Springfield, Virginia 22151 (AD-718 847, MF $.95, HC $3.00)

Document Not Available from EDRS.

Descriptors—Behavioral Science Research, *Computer Oriented Programs, Display Systems, Educational Environment, Educational Games, Hypothesis Testing, Input Output Devices, Interaction Mathematical Models, Mathematics Instruction, *Simulation, Statistical Analysis, Typewriting

Identifiers—APL, A Programing Language, Cathode Ray Tube, CRT, Florida State University, IBM 1500 Instructional System, Statistical Simulation, STATSIM

Two educational computer simulations are described in this paper. One of the simulations is STATSIM, a series of exercises applicable to statistical instruction. The content of the other simulation is comprised of mathematical learning models. Student involvement, the interactive nature of the simula ns, and ter inal display of materi s are fea common both sim s- tions. learn tions o cr imil l a s

The terms selected for our first group, Cathode Ray Tube, are therefore:

Closely Related:

CATHODE RAY TUBE
CRT
DISPLAY SYSTEMS
INPUT OUTPUT DEVICES
ON LINE SYSTEMS
MAN MACHINE SYSTEMS
SCREENS (DISPLAYS)

Broader Terms (could be dropped if output too high)

COMPUTER ORIENTED PROGRAMS
COMPUTER SCIENCE
COMPUTERS
ELECTRONIC DATA PROCESSING

Let's take Text Editing now.

TEXT EDITING
FAMILY OF TERMS

ECONOMICS
CONSUMER ECONOMICS
HOME ECONOMICS EDUCATION
EDUCATIONAL ECONOMICS
HOME ECONOMICS
LABOR ECONOMICS
OCCUPATIONAL HOME ECONOMICS
RURAL ECONOMICS
HOME ECONOMICS SKILLS
HOME ECONOMICS TEACHERS
EDITING
EDITORIALS
EDUCABLE MENTALLY HANDICAPPED
EDUCATION

TEXTBOOK ASSIGNMENTS
TEXTBOOK BIAS
TEXTBOOK CONTENT
TEXTBOOK EVALUATION
TEXTBOOK PREPARATION
TEXTBOOK PUBLICATIONS
TEXTBOOK RESEARCH
TEXTBOOK SELECTION
TEXTBOOK STANDARDS
TEXTBOOKS
HISTORY TEXTBOOKS
MULTICULTURAL TEXTBOOKS
SUPPLEMENTARY TEXTBOOKS
TEXTILES INSTRUCTION
HORIZONTAL TEXTS
PROGRAMED TEXTS
VERTICAL TEXTS
TEXTUAL CRITICISM
THAI
THEATER ARTS

ERIC Thesaurus - Alphabetic Display

EDITING 080
SN TO MAKE SUITABLE FOR PUBLICATION OR FOR PUBLIC PRESENTATION BY SELECTING, EMENDING, REVISING, AND COMPILING
UF COPYEDITING
BT EVALUATION METHODS
RT CONTENT ANALYSIS
FILMS
JOURNALISM
LANGUAGE ARTS
LANGUAGE STYLES
NEWS MEDIA
PUBLICATIONS

TEXTUAL CRITICISM
BT LITERARY CRITICISM
RT ANALYTICAL CRITICISM
CHRONICLES
FORMAL CRITICISM
HISTORICAL CRITICISM
ITALIAN LITERATURE
LITERARY ANALYSIS
110 LITERARY CONVENTIONS
LITERARY GENRES
LITERATURE

Combined Descriptor/Identifier List

1 EDITH GREEN
1 EDITING
1 EDITING AS A WAY OF LIFE
3 EDITING PROCEDURES
27 EDITORIALS
1 EDL READING VERSATILITY TESTS

1 TEXAS, UNIVERSITY OF TEXAS AT AUST
3 TEXT BOOKS
1 TEXT HANDLING SYSTEMS
1 TEXT PAC SYSTEM
1 TEXT PROCESSING
2 TEXT SEARCHING
10 TEXTBOOK ASSIGNMENTS
96 TEXTBOOK BIAS

ERIC Thesaurus -
Rotated Descriptor Display

Terms Suggested by Cross-Reference Structure

CONTENT ANALYSIS 190
BT EVALUATION METHODS
RT COMMUNICATION (THOUGHT TRANSFER)
COURSE CONTENT
CRITICAL READING
DATA ANALYSIS
EDITING
ITEM ANALYSIS
LITERARY ANALYSIS
LITERARY CRITICISM
LITERATURE REVIEWS
TEXTBOOK CONTENT

LITERARY ANALYSIS 260
NT LITERARY DISCRIMINATION
BT EVALUATION METHODS
LITERATURE
RT ANALYTICAL CRITICISM
CHARACTERIZATION (LITERATURE)
COMEDY
COMPOSITION (LITERARY)
CONTENT ANALYSIS
CRITICAL READING
DRAMA
EPICS
FIFTEENTH CENTURY LITERATURE
FIGURATIVE LANGUAGE
FORMAL CRITICISM
FRENCH LITERATURE
GERMAN LITERATURE
HISTORICAL CRITICISM
LITERARY CONVENTIONS
LITERARY CRITICISM
LITERARY GENRES
LITERARY STYLES
LOCAL COLOR WRITING
MEDIEVAL ROMANCE

LITERARY CRITICISM 260
NT ANALYTICAL CRITICISM
ARISTOTELIAN CRITICISM
FORMAL CRITICISM
HISTORICAL CRITICISM
IMPRESSIONISTIC CRITICISM
LITERARY STYLES
MORAL CRITICISM
MYTHIC CRITICISM
PLATONIC CRITICISM
RHETORICAL CRITICISM
TEXTUAL CRITICISM
THEORETICAL CRITICISM
RT CHARACTERIZATION (LITERATURE)
CONTENT ANALYSIS
CRITICAL READING
DIALOGUE
DRAMATIC UNITIES
EIGHTEENTH CENTURY LITERATURE
ENGLISH NEOCLASSIC LITERARY PERIOD
EXISTENTIALISM
FIFTEENTH CENTURY LITERATURE

The terms selected for our second group, Text Editing, are therefore:

Closely Related:

EDITING
EDITING PROCEDURES
TEXT HANDLING SYSTEMS
TEXT PROCESSING
TEXT SEARCHING

More distantly related, but might have been used by indexers in absence of specific term:

CONTENT ANALYSIS
LITERARY ANALYSIS
LITERARY CRITICISM
TEXTUAL CRITICISM

The final search statement can be structured in several ways:

1. Simple Intersection of Two Groups

CRT or Cathode Ray Tube or Computer Oriented Programs or Computer Science or Computers or Display Systems or Electronic Data Processing or Input Output Devices or On Line Systems or Man Machine Systems or Screens (Displays)	AND	Content Analysis or Editing or Editing Procedures or Literary Analysis or Literary Criticism or Text Handling Systems or Text Processing or Text Searching or Textual Criticism

2. Absolute retrieval of documents indexed by highly specific, but low posted terms, with intersection of remaining terms in each group.

Cathode Ray Tude OR CRT OR Editing OR Text Handling Systems OR Text Processing OR Text Searching

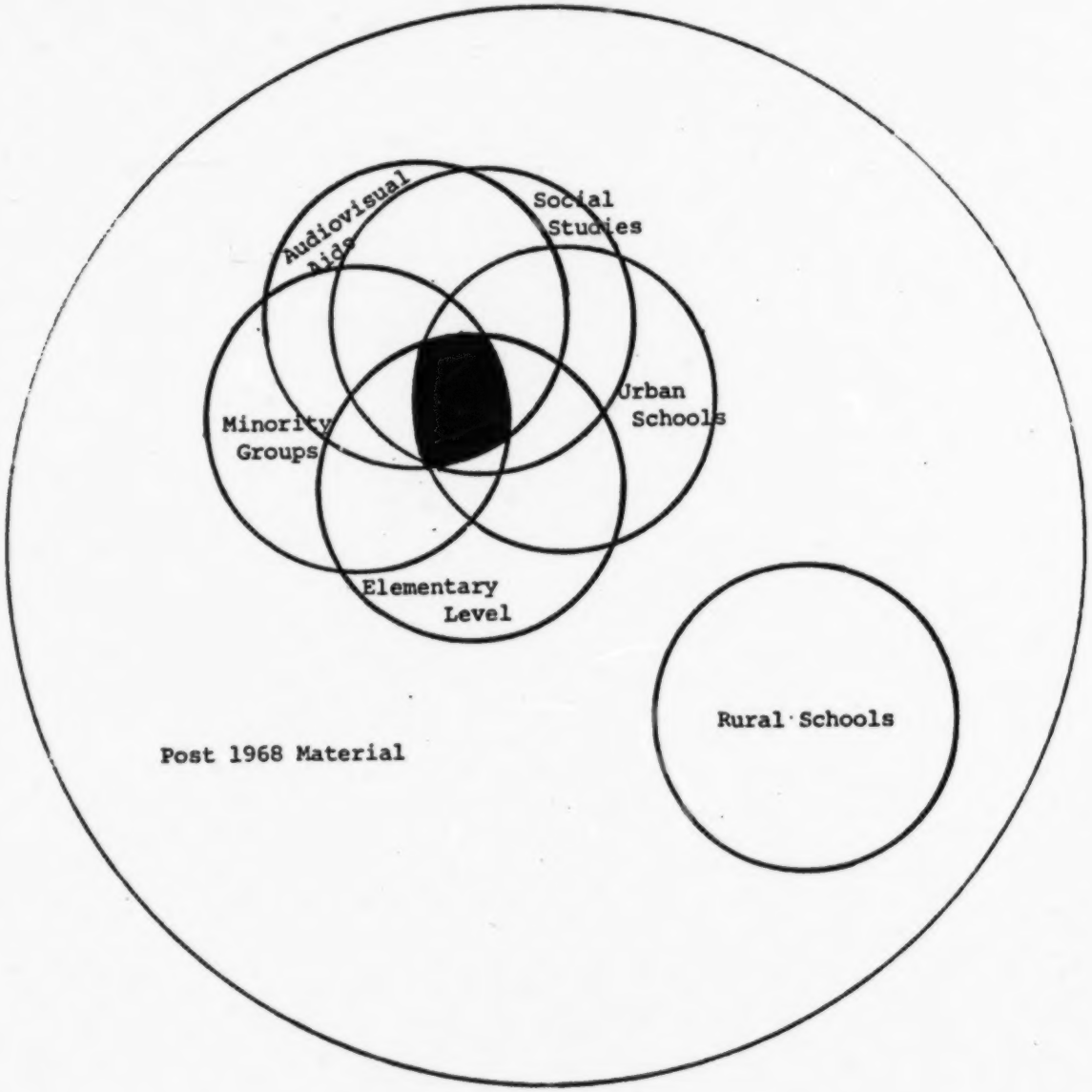
	OR	
Computer Oriented Programs or Computer Science or Computers or Display Systems or Electronic Data Processing or Input Output Devices or On Line Systems or Man Machine Systems or Screens (Displays)	AND	Content Analysis or Literary Analysis or Literary Criticism or Textual Criticism

In addition, the search involving the terms regarded as broader or peripheral could be handled separately in order to siphon off most of the "false drops" and low relevance material.

SEARCH #2

The use of audiovisual materials, instructional media, and innovative teaching techniques in teaching social studies to elementary level minority group children in urban school systems. Not interested in rural or small schools, or anything written before 1968.

This inquiry, on first inspection, seems to involve five major groups, a NOT function, and a date limitation, as shown below:



As the search strategy is considered, the searcher observes that it is better to NOT out the unwanted academic levels rather than to include "Elementary" level terms in an AND function, because academic levels are not always assigned by the indexers. However, including "Urban" terms in an AND group will effectively eliminate "Rural Schools", so a NOT function is not necessary to handle that particular restriction.

An effective way into the indexing vocabulary is to use the Rotated Descriptor Display for the following terms which appear in the inquiry:

Social Studies
Audiovisual
Media
Innovation
Minority
Urban
High Schools

The following groups begin to take shape:

FIRST GROUP		SECOND GROUP		THIRD GROUP
	AND		AND	
Social Studies		Audiovisual Aids		Minority Group Children
or		or		or
Social Studies Units		Audiovisual Instruction		Minority Groups
		or		
		Audiovisual Programs		
		or		
		Instructional Media		
		or		
		Instructional Innovation		

FOURTH GROUP		FIFTH GROUP
AND	BUT NOT	
Urban Schools		High Schools
		or
		High School Curriculum
		or
		High School Students
		or
		Junior High Schools
		or
		Junior High School Students

These terms can in turn be looked up in the Thesaurus, or the other tools, to add the following terms to the groups as follows:

SECOND GROUP		THIRD GROUP		FOURTH GROUP
	AND		AND	
Multimedia Instruction		Ethnic Groups		Urban Areas
		or		or
		Negroes		Urban Education
		or		or
		Negro Students		Urban Teaching

	FIFTH GROUP
BUT NOT	Secondary Schools

Check the postings to each term to get an idea of how many entries we are working with (e.g., URBAN SCHOOLS has over 800 terms, but since so many groups are being intersected, more "Urban" terms are picked up).

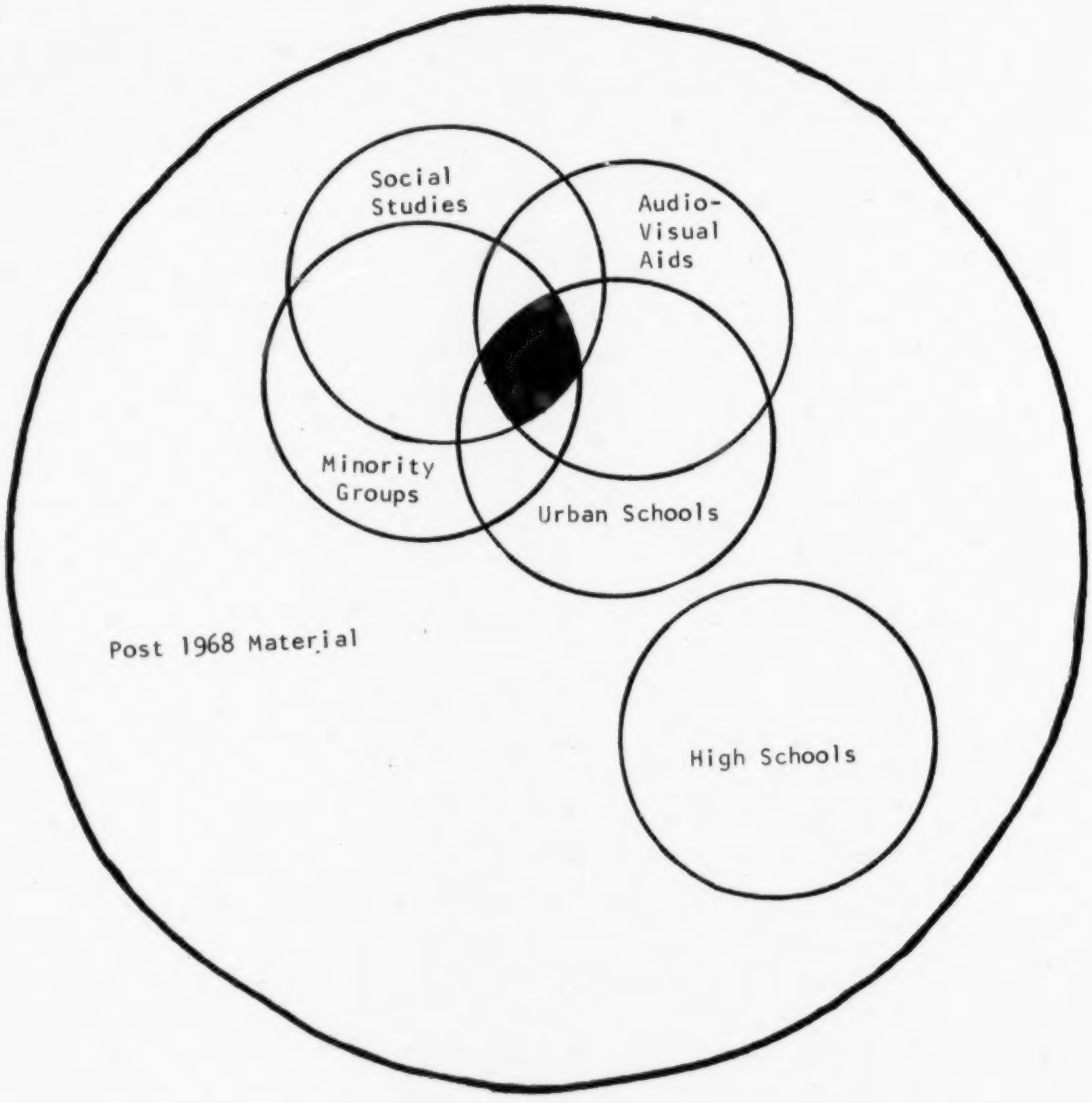
Limit the first group, "Social Studies", to major usage only since that is the requestor's prime concern, and since most teachers are not interested in wading through a lot of material.

The stipulation for no material written before 1968 must be handled according to the system capabilities---by ED number, by RIE issue, by publication date, or whatever your particular software allows.

The final logic and Venn diagram might be as shown below. Though the number of total terms is fairly large (23), do not be deceived; this is a very tight search (4 intersections) and could not result in a large output.

FIRST GROUP		SECOND GROUP		THIRD GROUP
	AND		AND	
Social Studies		Audiovisual Aids		Minority Group Children
or		or		or
Social Studies Instruction		Audiovisual Instruction		Minority Groups
		or		or
		Audiovisual Programs		Ethnic Groups
		or		or
		Instructional Innovation		Negro Students
		or		or
		Instructional Media		Negroes
		or		
		Multimedia Instruction		

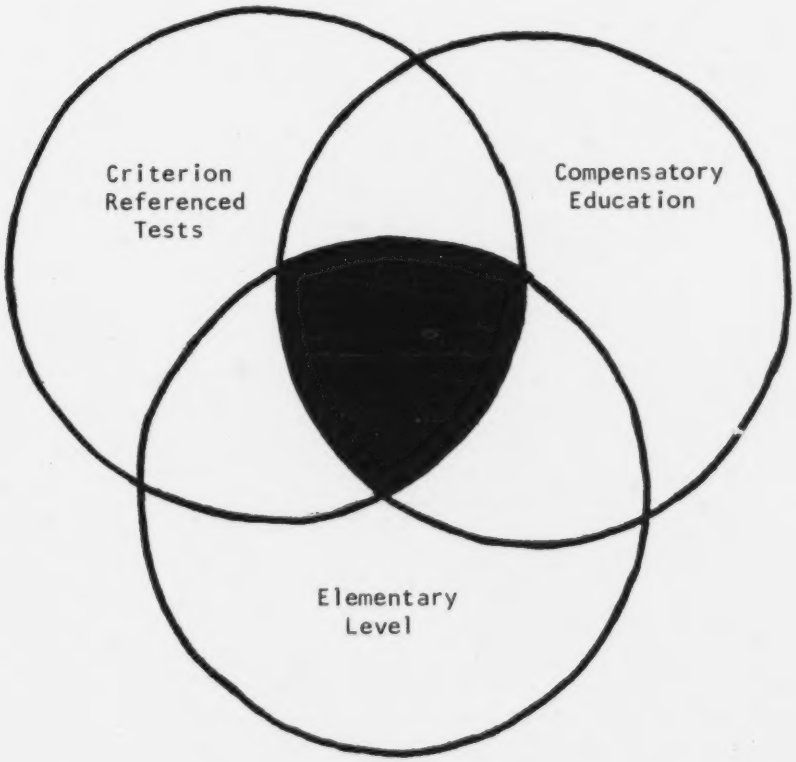
	FOURTH GROUP		FIFTH GROUP
AND		BUT NOT	
	Urban Areas		High School Curriculum
	or		or
	Urban Education		High School Students
	or		or
	Urban Schools		High Schools
	or		or
	Urban Teaching		Junior High School Students
			or
			Junior High Schools
			or
			Secondary Schools


Social Studies
Audio-Visual Aids
Minority Groups
Urban Schools
Post 1968 Material
High Schools

SEARCH #3

Research on the use of criterion reference testing in compensatory education programs, specifically ones funded under ESEA Title I. User thought California Test Bureau had done some work in this area. Elementary level; reading especially, but will accept other disciplines.

This is a three group search.



Let's suppose the searcher handling this request is not familiar with criterion tests. (If a telephone or interview request were involved, the searcher would have gotten as much information as possible from the requestor. If a letter of other remote means were involved, the searcher might well be on his own with the system.)

THESAURUS - ALPHABETIC DISPLAY

```
CRITERION MEASURES
 USE CRITERION REFERENCED TESTS

CRITERION REFERENCED TESTS                520
 SN  ANY TEST DESIGNED AND CONSTRUCTED
     ACCORDING TO EXPLICIT RULES LINKING
     AN INDIVIDUALS PERFORMANCE TO
     BEHAVIORAL REFERENTS
 UF  CRITERION MEASURES
     CRITERION TESTS
 BT  TESTS
 RT  ACHIEVEMENT TESTS
     DIAGNOSTIC TESTS
     GROUP TESTS
     ITEM BANKS
     MEASUREMENT TECHNIQUES
     NORM REFERENCED TESTS
     OBJECTIVE TESTS
     PROGNOSTIC TESTS
     STANDARDIZED TESTS
     TEST CONSTRUCTION

CRITERION TESTS
 USE CRITERION REFERENCED TESTS
```

THESAURUS - HIERARCHICAL DISPLAY

```
TESTS
.ACHIEVEMENT TESTS
..COLLEGE ENTRANCE EXAMINATIONS
..EQUIVALENCY TESTS
..LANGUAGE TESTS
..NATIONAL COMPETENCY TESTS
..PERFORMANCE TESTS
..READING TESTS
...INFORMAL READING INVENTORY
...READING READINESS TESTS
..SCIENCE TESTS
.APTITUDE TESTS
..INTEREST TESTS
..OCCUPATIONAL TESTS
.CLOZE PROCEDURE
.CREATIVITY TESTS
.CRITERION REFERENCED TESTS
.CULTURE FREE TESTS
.DIAGNOSTIC TESTS
.ESSAY TESTS
.GROUP TESTS
..GROUP INTELLIGENCE TESTS
..SCREENING TESTS
.INDIVIDUAL TESTS
.LISTENING TESTS
.NONVERBAL TESTS
.NORM REFERENCED TESTS
.OBJECTIVE TESTS
..MULTIPLE CHOICE TESTS
.PERCEPTION TESTS
.PHYSICAL EXAMINATIONS
..AUDITORY VISUAL TESTS
...AUDITORY TESTS
....AUDIOMETRIC TESTS
...VISION TESTS
..SPEECH TESTS
.PRESCHOOL TESTS
.PRETESTS
.PROBLEM SETS
.PROGNOSTIC TESTS
.PSYCHOLOGICAL TESTS
..ABSTRACTION TESTS
..ASSOCIATION TESTS
..COGNITIVE TESTS
..INTELLIGENCE TESTS
...GROUP INTELLIGENCE TESTS
..MENTAL TESTS
..PERSONALITY TESTS
...AFFECTIVE TESTS
...ATTITUDE TESTS
...IDENTIFICATION TESTS
...INTEREST TESTS
...MATURITY TESTS
...PROJECTIVE TESTS
...SELF CONCEPT TESTS
..SITUATIONAL TESTS
.SCHOOL READINESS TESTS
.STANDARDIZED TESTS
.TACTUAL VISUAL TESTS
.TIMED TESTS
.VERBAL TESTS
.VISUAL MEASURES
```

THESAURUS - GROUP DISPLAY

520 Tests

Devices or procedures for measuring ability, achievement, interest, etc., e.g., Achievement Tests, Aptitude Tests, Cognitive Tests, Interest Tests, Language Tests, Multiple Choice Tests, Problem Tests, Reading Tests, Talent Identification, Test Validity, etc.

520 Tests

ABSTRACTION TESTS
ACHIEVEMENT TESTS
AFFECTIVE TESTS
ANSWER KEYS
APTITUDE TESTS
ASSOCIATION TESTS
ATTITUDE TESTS
AUDIOMETRIC TESTS
AUDITORY TESTS
AUDITORY VISUAL TESTS
BIOGRAPHICAL INVENTORIES
CLOZE PROCEDURE
COGNITIVE TESTS
COLLEGE ENTRANCE EXAMINATIONS
CREATIVITY TESTS
CRITERION REFERENCED TESTS
CULTURE FREE TESTS
DIAGNOSTIC TESTS
EQUIVALENCY TESTS
ESSAY TESTS
GROUP INTELLIGENCE TESTS
GROUP TESTS
IDENTIFICATION TESTS
INDIVIDUAL TESTS
INFORMAL READING INVENTORY
INTELLIGENCE TESTS
INTEREST TESTS
LANGUAGE TESTS
LISTENING TESTS
MATURITY TESTS
MENTAL TESTS
MULTIPLE CHOICE TESTS
NATIONAL COMPETENCY TESTS
NONVERBAL TESTS
NORM REFERENCED TESTS
OBJECTIVE TESTS
OCCUPATIONAL TESTS
PERCEPTION TESTS
PERFORMANCE TESTS
PERSONALITY TESTS
PRESCHOOL TESTS
PRETESTS
PROBLEM SETS
PROGNOSTIC TESTS
PROJECTIVE TESTS
PSYCHOLOGICAL TESTS
PUZZLES
READING READINESS TESTS
READING TESTS
SCHOOL READINESS TESTS
SCIENCE TESTS
SCREENING TESTS
SELF CONCEPT TESTS
SITUATIONAL TESTS
SPEECH TESTS
STANDARDIZED TESTS
TACTUAL VISUAL TESTS
TALENT IDENTIFICATION
TEST CONSTRUCTION
TEST SELECTION
TEST VALIDITY
TESTING PROBLEMS
TESTING PROGRAMS
TESTS
TIMED TESTS
VERBAL TESTS
VISION TESTS

FROM: ERIC CLEARINGHOUSE SCOPE OF INTEREST MANUAL

TESTS, MEASUREMENT, AND EVALUATION (TM)

ADDRESS: ERIC Clearinghouse on Tests, Measurement, and Evaluation
Educational Testing Service
Rosedale Road
Princeton, New Jersey 08540

TELEPHONE: (609) 921-9000 X2691

ABSTRACT OF SCOPE:

Tests, scales, inventories, or other measurement devices or instruments; test development and construction; critical review of tests; measurement and evaluation procedures and techniques; applications and procedures of measurement or evaluation in educational projects or programs; comparative analysis of specific testing techniques.

APPLICABLE PHRASES AND TERMS (ALPHABETICALLY ARRANGED):

Aptitude Tests
Attitude Tests
Evaluation Procedures
Evaluation Techniques
Inventories
Measurement Procedures
Measurement Techniques
Scales
Tests

NOTE:

Documents concerned primarily with the procedures and techniques used in a project to evaluate, measure, or test certain variables (whatever the content, population, or level of the study itself may be) should be directed to ERIC/TM. If however, the interest is mainly on the subject matter, and evaluation plays only an incidental role, the document should be forwarded to the appropriate subject-oriented Clearinghouse.

FROM: ERIC CLEARINGHOUSE SCOPE OF INTEREST MANUAL

ERIC THESAURUS DESCRIPTORS

MOST COMMONLY USED IN INDEXING DOCUMENTS ON TESTS, MEASUREMENT, AND EVALUATION

Academic Achievement
Academic Performance
Achievement Needs
Attitude Tests
Behavioral Objectives
Classroom Observation Techniques
College Students
Correlation
Decision Making
Educational Improvement
Educational Objectives
Educational Research
Evaluation
Evaluation Criteria
Evaluation Methods
Evaluation Techniques
Factor Analysis
Item Analysis
Mathematical Models
Measurement
Measurement Instruments
Measurement Techniques
Models
Predictive Ability (Testing)
Predictive Measurement
Predictor Variables
Program Effectiveness
Program Evaluation
Questionnaires
Rating Scales
Research Methodology
Statistical Analysis
Student Attitudes
Student Evaluation
Test Construction
Test Interpretation
Test Reliability
Test Validity
Testing
Tests

With the help of the Thesaurus scope note, and any information provided by the requestor, we might decide to use the following descriptors:

Criterion Referenced Tests
Diagnostic Tests
Prognostic Tests

Check the postings to see what quantities we're working with. Also check the Identifier Usage Report for possible additional terms. The following identifiers might be added to complete the first group:

Criterion Referenced Measurement
Criterion Tests
Dignostic Reading Program
Diagnostic Reading Tests
Diagnostic Reading Tests
Diagnostic Tests (Education)
Prognostic Tests (Education)

The second group of terms can largely be located in the Rotated and Alphabetic Displays of the Thesaurus, and in the Identifer Usage Report.

California Test Bureau
Compensatory Education
Compensatory Education Programs
Elementary Secondary Education Act Title I
Elementary Secondary Education Act Title I Program
ESEA Title I
ESEA Title I
ESEA Title I Programs
Reading Programs
Remedial Instruction
Remedial Programs
Remedial Reading
Remedial Reading Programs

Because the postings in the main group, Criterion Referenced Tests, are not very high, it is again best to NOT out the unwanted levels, or ignore the levels entirely, rather than use elementary level terms in an AND function. This forms the third group.

The final search statement might look like this:

FIRST GROUP		SECOND GROUP
	AND	
Criterion Referenced Measurement		California Test Bureau
or		or
Criterion Referenced Tests		Compensatory Education
or		or
Criterion Tests		Compensatory Education Program
or		or
Diagnostic Reading Program		Elementary Secondary Education Act Title I
or		or
Diagnostic Reading Test		Elementary Secondary Education Act Title I Program
or		or
Diagnostic Reading Tests		ESEA Title I
or		or
Dianostic Tests		ESEA Title I
or		or
Diagnostic Tests (Education		ESEA Title I Programs
or		or
Prognostic Tests		Reading Programs
or		or
Prognostic Tests (Education)		Remedial Instruction
		or
		Remedial Programs
		or
		Remedial Reading
		or
		Remedial Reading Programs

	THIRD GROUP
BUT NOT	High School Curriculum
	High School Students
	High Schools
	Junior High School Students
	Junior High Schools

If the main concern were judged to be the use of criterion referenced tests in compensatory education, with the elementary level restriction not a strong one, a simple two level search should be run, dropping the negated third group altogether, particularly in light of the relatively low postings to the terms in Group 1.

However, the user definitely wants only elementary level material, a three-way intersection could be used, substituting a postive third group of "elementary" terms for the negated non-elementary term group.

NEW THIRD GROUP TO BE ANDed WITH FIRST TWO GROUPS

Elementary Education
or
Elementary Grades
or
Elementary School Curriculum
or
Elementary Schools
or
Elementary School Students
or
Grade 1
or
Grade 2
or
Grade 3
or
Grade 4
or
Grade 5
or
Grade 6

The following example represents an attempt to put too many restrictions on this search. If anything emerged, it would be the "perfect hit", but it is much more likely to result in no hits.

FIRST GROUP

Criterion Referenced Measurement
or
Criterion Referenced Tests
or
Criterion Tests
or
Diagnostic Reading Program
or
Diagnostic Reading Test
or
Diagnostic Reading Tests
or
Diagnostic Reading Tests
or
Diagnostic Tests
or
Diagnostic Tests (Education)
or
Prognostic Tests
or
Prognostic Tests (Education)

AND

SECOND GROUP

Compensatory Education
or
Compensatory Education Programs

AND

THIRD GROUP

ESEA Title I
or
ESEA Title I
or
ESEA Title I Programs
or
Elementary Secondary Education Act Title I
or
Elementary Secondary Education Act Title I Program

AND

FOURTH GROUP

Reading Programs
or
Remedial Reading
or
Remedial Reading Programs
or

AND

FIFTH GROUP

Elementary Education
or
Elementary Grades
or
Elementary School Curriculum
or
Elementary School Students
or
Elementary Schools
or
Grade 1
or
Grade 2
or
Grade 3
or
Grade 4
or
Grade 5
or
Grade 6

This search obviously has several possible ways it can be handled. The search negotiation process would hopefully help the searcher determine the best strategy to fit the user's needs. Also of importance, however, would be the emperical results achieved. The multi-level intersections, while sophisticated, may be too limiting to be practical. If early attempts fail to yield any (or sufficient) hits, the searcher could easily be forced back on a "coarser sieve" (using fewer intersections and based on the basic terms in Group 1) in order to find more material for the user to peruse.

IV. OUTPUT PHASE

NO HITS - WHAT TO DO?

Let us assume that a search has been made and that there has been no machine failure involved. In other words, the search was submitted and processed by the computer, but no file records were found to meet the stated search specifications. What should be done next?

The following is one checklist that might be followed:

1. Was There an Error in Coding?

   There may be something technically wrong with the way the question was asked. This is sometimes referred to as an error in "syntax", the syntax referred to being the particular conventions of the search program as to how queries should be translated into symbols, i.e., coded. For example, the number of parentheses may incorrectly be odd instead of correctly even. What this amounts to is that the searcher didn't really ask the question he wanted to ask.

2. Were There Typographical Errors?

   Are all the index terms involved in the search spelled correctly? Unless the searcher has the advantage of a system that validates search terms against the Thesaurus, he may unknowingly have specified EDUCATON rather than EDUCATION. The computer searches for exactly what was asked for and, of course, finds nothing. In this same category are problems involving incorrect use of blanks. A blank space is a character like any other as far as the computer is concerned. If you put two blanks (or no blanks) between the words of a multi-word Descriptor, than the computer, being totally literal, looks for exactly that. As far as garnering hits is concerned, you might as well have written the Descriptor backwards.

3. Was the Logic Too Restrictive?

   It doesn't take much experience in searching before you discover that intersections (AND logic) drastically cuts down on the number of hits. In most systems A AND B is a format often utilized; A AND B AND C will rarely result in a large number of hits; A AND B AND C AND D will almost always result in no hits. (These generalizations depend, of course, on both the number of terms assigned on the average to each record, and on the total records in the file, but in most bibliographic search systems they will hold true). Check your logic to see how many intersections are involved.

4. Did You Check the Term Usage Statistics?

   Terms with very low usages should generally not be intersected with other terms as the probability of a hit will usually be low. It will usually be preferable to simply ask for all the usages of an infrequently used term (OR logic). This approach does not create excessive hits and does not run the danger of no hits.

5. Have You Selected Alternative Approached to the Search Topic?

Perhaps you have gone down only one trail (and that a dead-end) to get at what you want. There may be other trails, other Descriptors, quite close in meaning to the ones you selected, that the indexers have preferred to use in dealing with your topic. For example, you may have used SUMMER SCIENCE PROGRAMS, but neglected to also use SUMMER PROGRAMS intersected with various terms beginning with the word SCIENCE.

6. How Have Documents Similar To The Kind You Are Seeking Been Indexed?

Perhaps you haven't found the trail at all yet. One way to get started is to examine a known hit to see how it was indexed. You may pick up insights as to indexer approach that did not occur to you when contemplating the problem independently.

7. Is There Likely To Be Material on This Topic in the Data Base?

Perhaps the no hits situation is to be expected. The data base may simply not be likely to have material on the topic requested. For example, COMPUTER MEMORIES, NATURAL CHILDBIRTH, etc., are not going to be well represented in the ERIC data base, if at all.

RECALL AND RELEVANCE

Recall and Relevance (Precision) are twin concepts that have been developed in order to attempt to measure and evaluate the quality of searches.

Relevance, or Precision, as it has come to be called more and more, is a measure of whether the items received as output are relevant to the original inquiry. The decision as to whether an item is relevant can obviously be made by several people: the searcher, the user, a panel of judges or experts, etc. When discussing relevance it is essential to state who is making this decision.

Recall is a measure of how many of the relevant items in the file being searched were found. To what extent was the search comprehensive, did it exhaust the possibilities in the file? Was a lot of material left behind that the user would have wanted?

In order to better explain these two measures, let us construct a hypothetical situation:

Size of Total File	100,000 items
Number of References in File Which are Relevant to Inquiry A	100 items
Number of References Retrieved by Actual Search	80 items
Number of References Retrieved Which Are Judged to be Relevant to Inquiry A	60 items

Recall is defined as the following ratio:

$$\frac{\text{Number of Relevant References Retrieved}}{\text{Number of Relevant References in File}} = \frac{60}{100} = 60\%$$

Relevance is defined as the following ratio:

$$\frac{\text{Number of Relevant References Retrieved}}{\text{Total Number of Reference Retrieved}} = \frac{60}{100} = 75\%$$

Many studies have shown that there is an inverse relationship between these two measures. In other words, in order to capture the remaining relevant references (that were missed the first time) it is necessary to "cast the net" so wide that a number of irrelevant references are also retrieved. Imagine the following situation, for example:

	IDEAL RETRIEVAL	ACTUAL RETRIEVAL
Number of References in File Which Are Relevant	100	100
Number of References Retrieved	100	400
Number of Retrieval References Which Are Judged Relevant	100	100

In this example, the Recall ratio has risen to 100%, but Relevance has dropped to $\frac{100}{400} = 25\%$.

Conversely, any attempt to push the Relevance ratio up tightens the search and inevitably sends the Recall ratio down.

Experience suggests that a stable balance of about 65-80% Recall and 65-80% Relevance is about the best that a system can achieve. Figure 1 depicts how these two measures relate to one another.

It must be kept in mind that the negotiation with the user can often determine whether the searcher should strive for high Relevance or high Recall. In the former instance, the user loses the opportunity to make unexpectedly valuable "finds" among material which is partly related to his topic. In the latter instance, the user is being asked to accept (perhaps pay for) a heavy proportion of marginal material in order to cover his topic comprehensively.

As a general rule, however, it is advisable to err on the side of achieving Relevance, rather than the reverse. The reasons for this are:

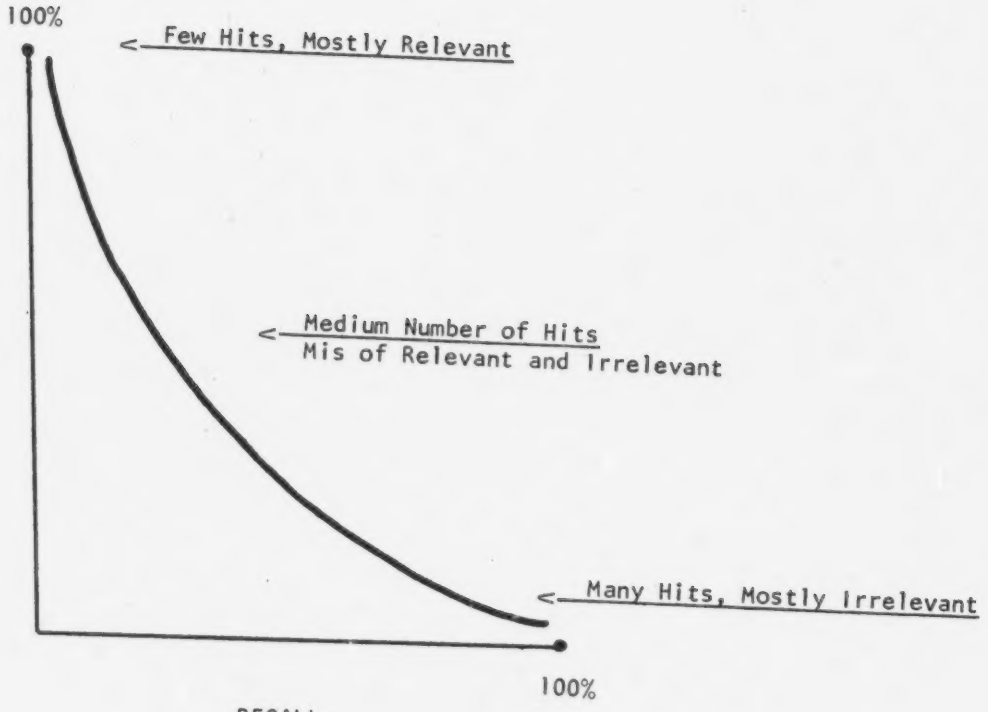
1. The user will not be expecting to use the computer as a browsing device. The usual stereotype of the computer will lead him to think of it as a fast and accurate method of receiving precisely the information "asked for". If the user receives a lot he did not "ask for", he will begin to question not so much the machine as the human operator doing the search.

2. Someone is, of course, paying for any excess retrieval; if not the patron, then perhaps the reference center.

In some situations the search strategy with respect to Relevance and Recall may be a matter of common sense policy based on factors other than the user. For example, in NASA's early days, when the file was small, and search reports were few, there was an editorial step in which output was examined and winnowed before transmittal. The policy during this period was to cast a wide net, i.e., high Recall. Later, when the file had grown in size and the number of search requests was large, the editorial step was dispensed with for economic reasons and the policy was to aim for high Relevance and immediate unedited transmittal of output.

RELEVANCE

(As relevance gets closer to 100%, you leave more and more items in the file; recall gets lower and lower, but never gets down to 0%.)



RECALL

(As you get everything in the file, relevance gets lower and lower, but never gets to 0%.)

OUTPUT VOLUMES. WHAT IS TOO LITTLE? WHAT IS TOO MUCH?

The answer to the question posed by this title will almost always depend on the user who asked the original question. I have personally seen real life situations where: (1) the user was hoping for 0 hits in order to verify that no one else was working on the topic he hoped to enter; (2) the user wanted about 5,000 hits in order to prove that a large government program of several years duration had resulted in significant volumes of research reports and other documentation.

It is very definitely a parameter that should be gotten from the user during the negotiation process. More often than not, if the searcher has a general idea of the user's anticipated or desired volume, he can control to meet this volume goal.

In the average search, however, it should be kept in mind that as the volume mounts it begins to approach a point where it will exceed the ability of the user to encompass it, to comprehend it, to make good use of it, even to read the titles of each item output. This upper limit will vary somewhat for each user because each user has a different threshold, a different ability to handle large output volumes. My own experience would set this upper limit at around 200 hits. I try to stay under 200 hits unless the user has specifically indicated that a comprehensive search is desired.

Sometimes, to avoid excessive output volumes, search systems will have a built in "hit limit" restricting output to some arbitrary number, e.g., 250, 400, 500, etc. The purpose of the "hit limit" is both to avoid inundating the user and also as insurance against a faultily constructed search that would otherwise "dump the file". "Hit limits" should only be used when the output emerges in reverse chronological sort, i.e., latest first. Otherwise the items that are over the limit and therefore dropped would be the latest and most up-to-date material. It is preferable to exclude the oldest hits, not the newest hits.

It is rare that a user will complain about too few hits, if they are genuinely relevant. The fewer hits there are, the less work the user has in reviewing them. For this he is perhaps unconsciously grateful. Low volume output can be a problem, however, if the hits are of marginal interest.

The best solution to this problem is to get the maximum amount of information from the user as to his problem, his application, and the end use of the search output. If the user wants a few items of high relevance for immediate use, the searcher's strategy and approach would be quite different than if the user wanted a comprehensive search of the file for the benefit of an extended state-of-the-art review.

APPENDIX A

ERIC VOCABULARY IMPROVEMENT PROGRAM

I. INTRODUCTION

A. Background

Establishing and maintaining, with limited resources, an indexing vocabulary for a system which has a subject field as broad as that of ERIC and which, in addition, is decentralized on a subject basis, has presented a number of unique problems. At the inception of ERIC, for example, an effort such as Project Lex (then in progress) was out of the question because the expense was more than could be supported or justified. The decision was made to let the documents being indexed determine the vocabulary.

To avoid overloading the vocabulary with seldom used, highly specific terms (such as personal names, test names, geographic locations, etc.), indexing was divided into two types: Descriptors, which would be included in a controlled hierarchically structured vocabulary (Thesaurus); and Identifiers, which would be uncontrolled, and unstructured, but which would permit use of specific indexing for precise retrieval.

The Descriptors which had been used for indexing the Disadvantaged Collection[1] in mid-1966 were chosen as a core vocabulary upon which the ERIC Thesaurus could be built. A Descriptor Justification Form (DJF) was designed to permit entry of new terms with possible synonyms (UF), broader terms (BT), narrower terms (NT), and related terms (RT). Provisions were also made for entering scope notes, a descriptor group identification code, and justification for the term selection, including authority citations. A set of rules[2] was published, and procedures for submitting candidate terms were established.[3]

Briefly, the procedures call for submittal of a DJF for a candidate term when---and only when---the term is required for indexing a document in hand. The DJF is prepared by the indexer from one of the ERIC Clearinghouses, reviewed by Clearinghouse supervision, and forwarded with a copy of the document input resume to the ERIC Processing and Reference Facility, which is the central switching point for the network. At the Facility, the DJF is reviewed and edited by a lexicographer for consistency, avoidance of proliferation, clarity, and conformance with rules and guidelines. Further review is imposed at the discretion of Central ERIC (National Institute of Education).

On the whole, this procedure has been quite successful. Thesaurus growth, which was quite rapid during the early years, has slowed markedly in the last several years, and is now relatively stable at around 5,000 main (postable) terms. However, the vocabulary is by no means perfect. With up to 20 different organizations scattered across the country indexing documents and submitting candidate terms, with the pressures of meeting publication deadlines working against extensive research

---

1. Catalog of Selected Documents on the Disadvantaged---Subject Index, ED 070 485; Number & Author Index, ED 070 484.

2. Rules for Thesaurus Preparation, OE-12047 (Superintendent of Documents, Washington, D.C., Sept. 1969).

3. ERIC Processing Manual, Thesaurus Section (ERIC Processing & Reference Facility).

and coordination, and with substantial turnover in some Clearinghouses, it was inevitable that some mistakes would occur and that some less-than-optimal decisions would be made. Over the years, a number of shortcomings in the vocabulary have developed:

- Poor, incomplete, or invalid hierarchies;
- Synonymy - Two or more terms which, for the purposes of ERIC indexing and retrieval, can be considered synonyms, e.g., HEREDITY and GENETICS;
- Poor word choices, e.g., PUBLICIZE rather than PUBLICITY;
- Misspellings, e.g., PARODOX for PARADOX;
- Ambiguity, e.g., prior to the introduction of PROGRAMING (BROADCAST) in 1971, the term PROGRAMING had been applied to both computer programming and broadcast programming;
- Low postings, e.g., HORIZONTAL TEXTS and VERTICAL TEXTS with one posting each from the 1966 Disadvantaged Collection;
- Scattering in the Identifier file, e.g., 17 variations in entries for Title III of the Elementary and Secondary Education Act.

Unfortunately, correcting most of these shortcomings is not accomplished simply, particularly when they have had time to "set". In the case of hierarchical defects, making a change is mechanically relatively easy, since only one or two DJF's are usually required. The problem arises in making sure that the change is in fact a correction---i.e., that the new structure is better than the old, and that there are no unwanted side effects. On the other hand, the other deficiencies are intellectually rather simple---you pick the preferred term and eliminate the non-preferred one(s). However, the implementation mechanics are complex and cumbersome. The ERIC software, which was designed to insure synchronization between the Resume Master Data Set (linear file) and the Satellite Master Data Sets (inverted files), will not permit the deletion of a term from the Thesaurus so long as there are documents posted to (indexed by) that term. Until recently, in order to delete a Thesaurus term, it was necessary to prepare a separate transaction for each document indexed by that term to delete the term from the Resume Master Data Set, and it would then be deleted from the Satellite Master Data Set by the system. At the same time, if you wanted to avoid an intolerable loss of information, a second set of transactions had to be prepared, replacing the deleted term with the preferred one. Since the median posting density of Thesaurus terms is about 50 documents per term, about 100 transactions would typically be required to accomplish each change. Obviously, not many changes could be made under those conditions.

Recently, the ERIC Facility completed and tested software which will permit changes of this type---to either Descriptors or Identifiers---with a single transaction which deletes a term and transfers its postings (if desired) to another term (or terms). With this added capability, the ERIC network is now in a position to make all of the changes required to develop its vocabulary into an optimal tool for indexing and retrieval. This, however, is not a task which can be performed in a vacuum by an individual or even by a single group. Above all, the vocabulary must be responsive to the needs of the system it serves, and this means primarily the people of all components, most assuredly including the users of the system outputs.

B. Vocabulary Improvement Program

The ERIC Vocabulary Improvement Program must be an integrated operation. A particular emphasis is given to system-wide participation, and vocabulary change recommendations of any kind are solicited from all components and users of the system. These include recommended changes in vocabulary conventions, vocabulary structure, and the basic terminology. A multi-faceted approach has been chosen to implement the program. There are three major facets, and these can---and must---be implemented in somewhat different fashions.

- Descriptor Cross-Reference Changes. These are changes in the BT, NT, and RT references in the Thesaurus itself and do not affect directly the question of which documents are indexed by which term(s). Consequently, these changes have little if any impact on the existing data base. Further, evaluation of cross-reference changes requires that they be viewed in the context of the surrounding "terminology terrain" which requires display of---at the very least---a significant portion of the Thesaurus, if not its entirety. Full-scale coordination of cross-reference changes among ERIC users is not anticipated as such activity would prove burdensome in terms of dissemination costs and evaluation time.

- Descriptor Changes. These are changes to the Thesaurus indexing as it exists in the data base, where a given Descriptor is removed from the file and its postings are transferred to one or more existing Descriptors or to a new Descriptor added to the file for this purpose. Since these changes have an immediate and significant impact on users of the file, as well as on day-to-day operations of the ERIC network, the widest practical coordination base is desired.

- Identifier Changes. A program to detect and correct Identifier variations has been implemented. The data base is being corrected via transfer-and-delete operations. Since the Identifiers are by design unstructured and uncontrolled, full user coordination at the level required for Descriptor changes is not anticipated.

The second and third facets encompass actual changes to the indexing terminology and are the subject of Section II which follows.

II. TERM CHANGE PROCEDURES

Differences between Descriptors (Thesaurus terms) and Identifiers dictate somewhat different procedures for the implementation of changes. Descriptor changes, which are more closely controlled, are discussed first.

A. Descriptor (Thesaurus) Changes

Changes in Descriptors or Thesaurus terms will be based largely upon usage in the data base and upon the detection and correction of situations of postable synonyms appearing in the vocabulary. Also, obvious misspellings and word-form corrections will be required in some instances. The following paragraphs discuss Descriptor changes based on usage data, Descriptor changes based on the elimination of synonyms, and the proposed procedures to be used in actually accomplishing changes.

1. Descriptor Editing Based on Usage.

Descriptors that are posted very heavily (over 1000 postings) should be examined for their utility. Some of these Descriptors may be quite valid (e.g., TEACHER EDUCATION) and very reflective of the emphases in the data base. However, the heavy postings on some others may indicate that they are too general to be useful in either announcement media or manipulative retrieval (e.g., EDUCATIONAL PROGRAMS). It may be that Descriptors in this latter case should be either: (1) names of Descriptor Groups; (2) "array" terms, each with a scope note cautioning against its use in indexing and retrieval and with cross-references to more specific Descriptors constituting the "tops" of appropriate generic families; or (3) provided with delimiting scope notes to avoid ambiguous usage in the future.

Descriptors used too infrequently tend to "clutter", unnecessarily impeding easy use of the data base, the indexes, and the Thesaurus. Descriptors used less than about five (5) times should be examined for possible removal from the active vocabulary, except for relatively recent additions to the vocabulary. "Old," low-usage descriptors should be either: (1) converted to nonpostable terms, with USE references inserted and index postings transferred to the referred-to-Descriptors; (2) deleted from the Thesaurus, but with postings transferred to selected Descriptors; or very occasionally, (3) deleted entirely from the Thesaurus, with postings also deleted from the data base.

2. Elimination of Postable Synonyms.

If postable synonyms exist in the Thesaurus, some documents will be indexed by one such synonym and some by the other. Retrieval via one synonym will thus be incomplete. Such a condition is highly undesirable. Instances of postable synonyms must be detected, preferred versions selected, USE references to these preferred versions created, and data-base postings transferred from the nonpreferred term to the preferred Descriptor.

3. Thesaurus Change Procedure.

The flow chart of the Thesaurus Change Procedure is shown in Figure 1. While it is anticipated that many recommendations for change will originate from the day-to-day work of the Facility Lexicographer (e.g., with term cross-references), change recommendations are solicited from the entire ERIC network and all users of the system. All changes, whether from internal or external

THESAURUS CHANGE PROCEDURE

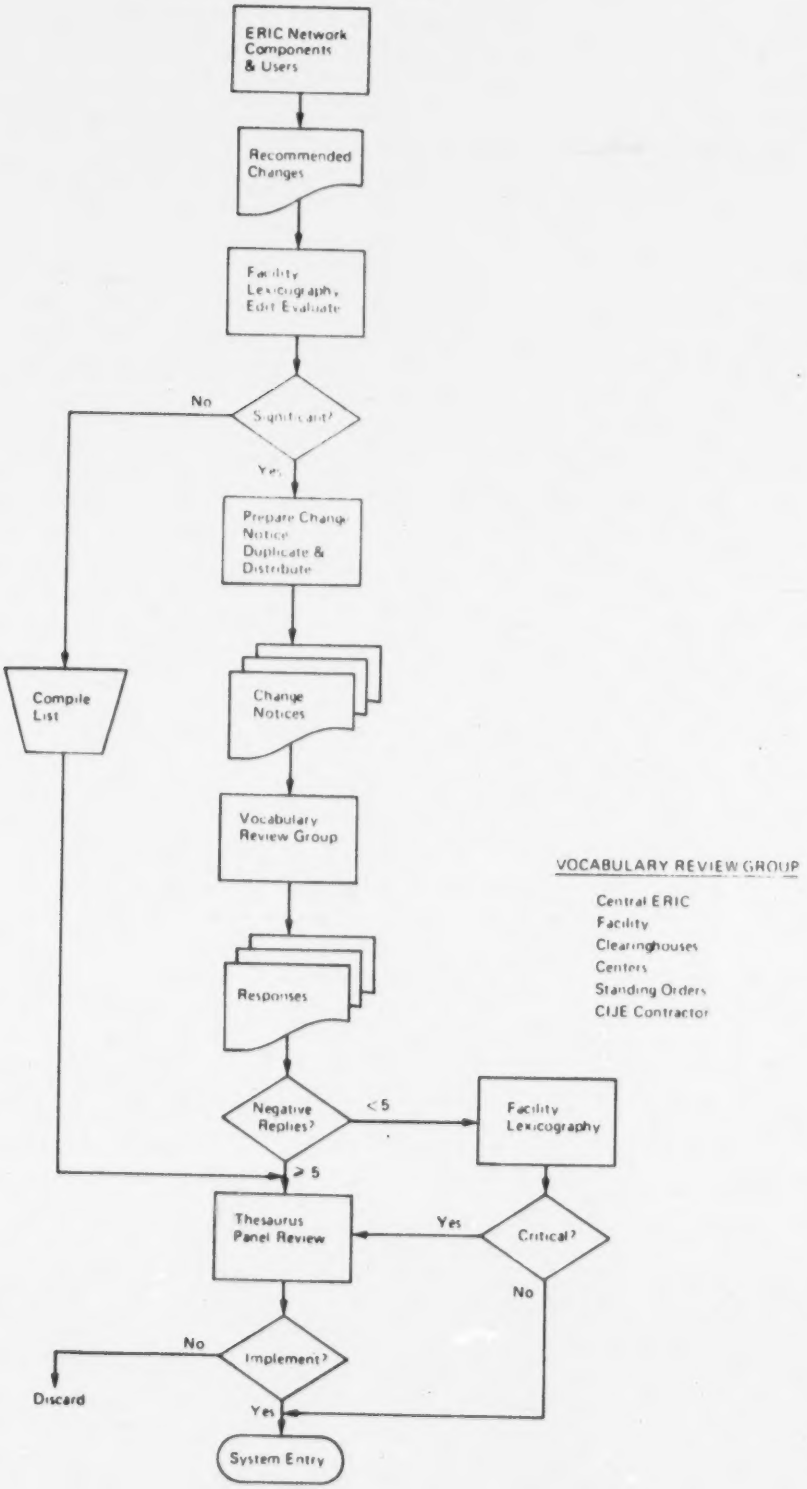


FIGURE 1.

sources, will be processed as shown in Figure 1.

a. Kinds of Changes

There are a number of different kinds of changes which can be made:

- Simple Merge---Used to eliminate synonyms and to post low-use terms to the next higher generic level.

  Examples: Transfer postings on HEREDITY to GENETICS

  Transfer postings on GIRLS CLUBS to YOUTH CLUBS

- Word Change---Used to correct misspellings and change word forms.

  Examples: Transfer postings on PARODOX to (new term) PARADOX

  Transfer postings on PUBLICIZE to (new term) PUBLICITY

- Multiple Merge---Used to eliminate multiple synonyms or to post several low-use terms to next higher generic level. May include word change.

  Examples: Transfer postings on MARKS and GRADES (REPORT) to GRADES (SCHOLASTIC)

  Transfer postings on QUICHE and YUCATEC to MAYAN LANGUAGES

  Transfer postings on HETEROPHORIA and HETEROTROPIA to (new term) STRABISMUS

- Term Split---Used to post low-use terms to two (or more) more general terms (not necessarily broader terms of the term in question), when transfer to the next higher generic level might result in significant information loss. The receiving terms can then be coordinated for searching to retain specificity.

  Examples: Transfer terms on FORESTRY OCCUPATIONS to FORESTRY and AGRICULTURAL OCCUPATIONS

  Transfer terms on OCEAN ENGINEERING to OCEANOLOGY and ENGINEERING

- Simple Delete---Used to remove terms which have been added to the Thesaurus erroneously, or which have proved to have no utility.

  Examples: Delete postings from SATELLITE LABORATORIES

  Delete postings from HORIZONTAL TEXTS

  Delete postings from VERTICAL TEXTS

The transfer-and-delete programs automatically generate transactions to purge (delete) terms from the Thesaurus when their postings are transferred. If a cross-reference is desired (e.g., HEREDITY Use GENETICS), this must be added separately. If a term deleted from the Thesaurus should be used as an Identifier, it is necessary (at the present time) to generate/add a separate Identifier transaction for each document indexed by the term to retain the postings.

b. Change Recommendations

(1) Information Required - In order for the change to be processed efficiently, the following information is required for each change proposed:

- Statement of the Desired Change---Simple, imperative sentences like those used in the examples above are preferred.

- The Number of Postings---Required for each term involved in the change, this information may be obtained from the publication ERIC Descriptor and Identifier Usage Report. The date (month/year) of the postings count should be noted. For changes other than synonyms and word form changes, the accession number of the last known document indexed by each term in also desired. The latest accession number will indicate the timeliness of the terminology in question.

- Reason for Change---e.g., eliminate synonyms, correct spelling, etc.

- Justification for the Change---Unless the change is a correction of an obvious error, such as a misspelling (PARODOX/PARADOX), justification for the change must be supplied. Authorities for definitions should be indicated. Generally, in the case of synonyms, postings will be transferred to the term with the larger number of postings. If a given recommendation is to reverse this practice, the reasons for doing so must be explicit to justify the added expense. The timeliness of terminology should be examined before recommending the transfer of postings to a higher generic level, or a simple deletion. Low use is not per se sufficient reason for deletion; a certain amount of time has to be allowed for a new term to build up postings.

(2) Submittal of Recommendations - Change recommendations do not have to be in any particular format, so long as the required information is included. Recommendations should be addressed to:

ERIC Processing and Reference Facility
ATTN: Lexicographer
4833 Rugby Avenue, Suite 303
Bethesda, Maryland 20014

c. Evaluation and Edit

The Lexicographer will evaluate incoming Change Recommendations and separate them into three categories as follows:

- o Category 1 - Significant, having major impact on indexing and/or potentially controversial.
- o Category 2 - Obviously necessary or useful, having minor impact and/or not likely to be contested.
- o Category 3 - Obviously trivial, contrary to rules, or insufficient justification or support for the change.

It is anticipated that virtually all of the Change Recommendations will fall into Category 1 and be processed through the coordination procedure described in the following paragraphs. However, a few Category 2 changes can be expected, and Category 3 changes, while not anticipated, are possible. The Lexicographer will record all Category 2 and 3 change proposals in a list-form report which will be reviewed by the Thesaurus Advisory Panel (See paragraph g. below).

d. Change Notice

The Lexicographer will prepare for each Category 1 Change Recommendation a Thesaurus Term Change Notice (Figure 2, Form EFF-21) completing sections 1 through 3. The forms will then be duplicated and distributed in two (2) copies to each member of the ERIC Vocabulary Review Group.

e. Vocabulary Review Group Responsibilities

Each member of the Vocabulary Review Group designates a responsible individual (Vocabulary Coordinator) to review all Change Notices, coordinating internally as desired. The membership of this group has been chosen to achieve the broadest possible coordination base consistent with efficient operation.[4]

The Vocabulary Coordinator will review each Change Notice as received, complete the RECOMMENDED ACTION section, sign the form, and return one (1) copy to the Lexicographer at the ERIC Facility within two (2) weeks of receipt. This deadline is established to avoid unwarranted delays

---

4. A written invitation to join the Vocabulary Review Group (from C.W. Hoover, Chief, ERIC) was distributed in late June 1973 to a total of approximately 60 organizations. A total of 35 organizations responded favorably to this invitation, indicating interest in the Vocabulary Improvement Program and designating individuals who would participate. These 35 organizations make up the existing Vocabulary Review Group; their composition includes 16 ERIC Clearinghouses, 10 university libraries, and 9 agencies of state education departments.

ERIC

THESAURUS TERM CHANGE NOTICE

Seven (A)

No. _____

1. PROPOSED CHANGE

2. IMPACT

a. POSTINGS BEFORE CHANGE

Term _____ Postings

b. POSTINGS AFTER CHANGE

Term _____ Postings

3. REASON FOR CHANGE (Include full justification, citing authorities for definition, usage, and treatment)

RECOMMENDED ACTION

☐ CONCUR ☐ NO INTEREST

☐ OBJECT (State reasons in full detail, including potential impact upon input or retrieval operations showing significant loss of information. Cite authorities as appropriate.)

Signed:

Vocabulary Coordinator _____ Organization _____

RETURN PRIOR TO _____ To: ERIC Processing and Reference Facility
ATTN: Lexicographer
4833 Rugby Avenue, Suite 303
Bethesda, Maryland 20014

EFF 21 (8 72)

FIGURE 2.

or excessive follow-up. Failure to respond within the time limit established will be treated as an indication of concurrence or lack of interest in the change.

f. Tabulation of Responses

The Facility Lexicographer will tabulate responses to each Change Notice as they are received. After the cut-off date, the objections to each change will be counted.

- Significant Objections---If there are five (5) or more objections, the Change Notice will be set aside for review by the Thesaurus Advisory Panel at its next session.

- None/Few Objections---If there are fewer than five (5) objections, the Lexicographer will examine the objections received to determine whether or not the change is likely to have a critical impact on an objector's operations. If so, the Change Notice will be set aside for review by the Thesaurus Advisory Panel. If not, the change will be entered into the system.

g. Thesaurus Advisory Panel Review

(1) Schedule - The Thesaurus Advisory Panel[5] will confer at least quarterly, usually during the 3d week of January, April, July, and October. Additional meetings may be scheduled, as necessary. This schedule is timed to permit the decisions of the Panel to be incorporated into the file prior to release of cumulative indexes, the quarterly Thesaurus updates, the quarterly ERICTAPE updates, and the annual issuance of the ERIC Descriptor and Identifier Usage Report.

(2) Agenda - Depending upon the material available for consideration, the Panel will take action in the following areas:

- Review, approve, disapprove, or modify changes in Group Codes and cross-reference structure.

- Examine the list-form report of Category 2 and Category 3 Change Recommendations; confirm or reverse decisions, or re-classify items to Category 1 for full coordination.

- Consider all Change Notices to which five (5) or more objections have been received---as well as those judged critical; examine pros and cons of each change in the light of total system needs, and determine disposition.

- Discuss other vocabulary-related matters, including such future plans and programs as: rules changes or clarifications, format changes, and publications changes.

---

5. The Thesaurus Advisory Panel includes a Chairperson (Central ERIC), a lexicographer (ERIC Facility), an ERIC Clearinghouse representative, and 5-7 other members selected from public and private agencies. As of this writing, the final composition of this group has not been determined.

h. Implementation and Dissemination

Immediately after each Panel meeting, approved changes will be implemented, using the transfer-and-delete software, so that changes will be incorporated in the next edition of the Thesaurus, RIE cumulative indexes, etc. In addition, lists of all changes will be incorporated in the next issues of ERIC Management Notes and Interchange.

B. Identifier Changes

1. Identifier Scattering.

ERIC Identifiers are essentially uncontrolled, and a great many synonyms have crept into the Identifier list over the years---e.g., different forms and/or abbreviations and/or syntactical variants of names of organizations. It is often impractical for the user, with this state of affairs, to insure that he has detected all variants of a particular Identifier. A program to detect and correct Identifier variations has been implemented. For each set of Identifier synonyms, a preferred version is being selected. The data base will be corrected via transfer-and-delete operations. Since Identifiers are by design neither structured nor controlled, full user coordination will not be required.

2. Sources and Information Required.

Identifier Change Recommendations (as with Descriptors) are solicited from all ERIC components and users. Change recommendations should generally conform to the pattern specified for Descriptors, except that justifications need not be as complete.

3. Review and Edit.

Identifier Change Recommendations will be reviewed and edited by the Facility Lexicographer and automatically assigned to either Category 2 or 3.

4. Coordination.

A list-form report of the Lexicographer's decisions on Identifier Change Recommendations will be submitted to the Thesaurus Advisory Panel along with the corresponding Descriptor (Thesaurus) Change Recommendations for confirmation, reversal, or re-classification. Re-Classified Identifier Change Recommendations will be fully coordinated in the same manner as Thesaurus Change Notices.

5. Implementation and Dissemination.

Implementation and dissemination will be accomplished in the same manner as Descriptor changes.

THESAURUS

ERIC

TERM CHANGE

NOTICE

No. 1

1. PROPOSED CHANGE Transfer postings on PERSONAL RELATIONSHIP to INTERPERSONAL RELATIONSHIP. Retain PERSONAL RELATIONSHIP as UF to INTERPERSONAL RELATIONSHIP.

2. IMPACT

a. POSTINGS BEFORE CHANGE (Dec '72, RIE)

Term			Postings
PERSONAL	RELATIONSHIP	(Major)	6
"	"	(Minor)	17
INTERPERSONAL	RELATIONSHIP	(Major)	139
"	"	(Minor)	275

b. POSTINGS AFTER CHANGE

Term			Postings
INTERPERSONAL	RELATIONSHIP	(Major)	145
"	"	(Minor)	292

3. REASON FOR CHANGE (Include full justification, citing authorities for definition, usage, and treatment)

Both PERSONAL RELATIONSHIP and INTERPERSONAL RELATIONSHIP are very old descriptors, dating back to the Phase I ERIC Thesaurus (pre-1968). Originally, the two terms were not cross-referenced, indicating that one (the second to be entered) was added without knowledge of the other; currently, INTERPERSONAL RELATIONSHIP is the broader term. PERSONAL RELATIONSHIP might conceivably be used to refer to a more basic or intimate relationship (especially between two people) than INTERPERSONAL RELATIONSHIP might imply. However, this distinction is unnecessary for an educational vocabulary. See "Interpersonal" and "Personal" in English & English's Comprehensive Dictionary of Psychological & Psychoanalytical Terms.

RECOMMENDED ACTION

☐ CONCUR ☐ NO INTEREST

OBJECT (State reasons in full detail, including potential impact upon input or retrieval operations showing significant loss of information. Cite authorities as appropriate.)

Signed

Vocabulary Coordinator _____ Organization _____

RETURN PRIOR TO October 12, 1973 To: ERIC Processing and Reference Facility
ATTN: Lexicographer
4833 Rugby Avenue, Suite 303
Bethesda, Maryland 20014

THESAURUS

ERIC

TERM CHANGE

NOTICE

No. 2

1. PROPOSED CHANGE

Change PLANNING (FACILITIES) with its UF "Facilities Planning," to FACILITY PLANNING.

2. IMPACT

a. POSTINGS BEFORE CHANGE

Term _____ Postings _____

b. POSTINGS AFTER CHANGE

Term _____ Postings _____

Not Applicable

3. REASON FOR CHANGE (Include full justification, citing authorities for definitions, usage, and treatment)

The term PLANNING (FACILITIES) does not conform to Item 1.1.3.1 of the ERIC Rules for Thesaurus Preparation. This rule states: "A parenthetical qualifier identifies any particular indexable meaning of a homograph. One of the reasons for restricting the use of parenthetical qualifiers to homographs is to preclude the use of inverted entries." The proposed term FACILITY PLANNING is in accord with this rule and is consistent with the rest of the "facility" terms in the Thesaurus.

RECOMMENDED ACTION

☐ CONCUR ☐ NO INTEREST

OBJECT (State reasons in full detail, including potential impact upon input or retrieval operations showing significant loss of information. Cite authorities as appropriate.)

Signed

Vocabulary Coordinator _____ Organization _____

RETURN PRIOR TO October 12, 1973 To: ERIC Processing and Reference Facility
ATTN: Lexicographer
4833 Rugby Avenue, Suite 303
Bethesda, Maryland 20014

ERIC

THESAURUS TERM CHANGE NOTICE

No. 3

1. PROPOSED CHANGE

Delete "Morals" as UF to ETHICS. Add MORALS as descriptor. Transfer postings on MORAL VALUES and ETHICAL VALUES to new term MORALS. Retain "Moral Values" and "Ethical Values" as UF's to MORALS.

2. IMPACT

a. POSTINGS BEFORE CHANGE (Dec '72, RIE)

Term	Postings
ETHICS	34
MORALS	0
ETHICAL VALUES	95
MORAL VALUES	130

b. POSTINGS AFTER CHANGE

Term	Postings
ETHICS	34
MORALS	225

3. REASON FOR CHANGE (Include full justification citing authorities for definition, usage, and treatment)

ETHICAL VALUES and MORAL VALUES are old Phase I Thesaurus terms that were never cross-referenced. They were probably entered as a result of free indexing, and without the benefit of lexicographic analysis. ETHICS--UF "Morals" was entered much later and structured using the LEX Thesaurus. The ambiguity and inconsistency among these terms could be eliminated with the above change and the addition of the following Scope Notes:

ETHICS....Study of the ideal in human character and conduct.
MORALS....Individual/group standards of conduct in terms of right or wrong, or actual conduct with reference to such standards.

See ETHICS and MORALS in English & English's Comprehensive Dictionary of Psychological & Psychoanalytical Terms.

RECOMMENDED ACTION

☐ CONCUR ☐ NO INTEREST

OBJECT (State reasons in full detail, including potential impact upon input or retrieval operations showing significant loss of information. Cite authorities as appropriate.)

Signed

Vocabulary Coordinator _____ Organization _____

RETURN PRIOR TO October 12, 1973

To: ERIC Processing and Reference Facility
ATTN: Lexicographer
4833 Rugby Avenue, Suite 303
Bethesda, Maryland 20014

EFF 21 (8-72)

ERIC

THESAURUS TERM CHANGE NOTICE

No. 4

1. PROPOSED CHANGE

Transfer postings on TEACHER EXPERIENCE to TEACHING EXPERIENCE. Retain TEACHER EXPERIENCE as UF to TEACHING EXPERIENCE, but drop the current UF "Professional Laboratory Experience."

2. IMPACT

a. POSTINGS BEFORE CHANGE (Dec '72, RIE)

Term		Postings
TEACHER EXPERIENCE	(Major)	36
" "	(Minor)	113
TEACHING EXPERIENCE	(Major)	23
" "	(Minor)	39

b. POSTINGS AFTER CHANGE

Term		Postings
TEACHING EXPERIENCE	(Major)	59
" "	(Minor)	152

3. REASON FOR CHANGE (Include full justification, citing authorities for definitions, usage, and treatment)

TEACHING EXPERIENCE was added to the Thesaurus in late 1969. It was believed that the existing descriptor TEACHER EXPERIENCE with its UF "Professional Laboratory Experience" was insufficient to express the idea of both preservice and inservice professional teaching experience taking place either in or out of a laboratory. This was a fallacy in that the UF should not have been construed as a delimiter. Thus, the new term was added in error. Some ambiguity will be eliminated by merging the postings and retaining the Scope Note for TEACHING EXPERIENCE--"Actual and simulated experience of preservice and inservice teachers" (Good's Dictionary of Education). Further ambiguity will be eliminated by adding a Scope Note to the descriptor TEACHER BACKGROUND. This Scope Note would simply state "Experience other than teaching."

RECOMMENDED ACTION

☐ CONCUR ☐ NO INTEREST

OBJECT (State reasons in full detail, including potential impact upon input or retrieval operations showing significant loss of information. Cite authorities as appropriate.)

Signed

Vocabulary Coordinator _____ Organization _____

RETURN PRIOR TO October 12, 1973

To: ERIC Processing and Reference Facility
ATTN: Lexicographer
4833 Rugby Avenue, Suite 303
Bethesda, Maryland 20014

FFF 21 8 72r

THESAURUS

ERIC

TERM CHANGE

NOTICE

No. 5

1. PROPOSED CHANGE

Transfer postings on HETEROPHORIA and HETEROTROPIA to new term STRABISMUS. Retain HETEROPHORIA and HETEROTROPIA as UF's to STRABISMUS.

2. IMPACT

a. POSTINGS BEFORE CHANGE (Dec '72, RIE)

Term	Postings
HETEROPHORIA	2
HETEROTROPIA	1

b. POSTINGS AFTER CHANGE

Term	Postings
STRABISMUS	3

3. REASON FOR CHANGE (Include full justification, citing authorities for definitions, usage, and treatment)

Two very specific terms (entered 2/68) with very low postings will be merged into a new, broader term. Both HETEROPHORIA and HETEROTROPIA refer to tendencies of the eyes to turn away from the position correct for binocular vision, but HETEROPHORIA is a "latent" imbalance or deviation in contrast to HETEROTROPIA or a "manifest" imbalance. STRABISMUS or "squint" is usually associated with HETEROTROPIA. It can, however, take a broader meaning (see "squint," Stedman's Medical Dictionary). As a new term, STRABISMUS will be scoped as follows:

STRABISMUS....Lack of coordination of eye muscles so that the two eyes do not focus on the same point.

In addition, the following current UF's to HETEROPHORIA and HETEROTROPIA will be dropped from the Thesaurus as they are not likely to appear in educational literature without reference to a more generic concept: "Cyclophoria," "Esophoria," "Esotropia," "Exophoria," "Exotropia," "Hyperphoria," "Hypertropia," "Hypophoria," and "Hypotropia." The UF's "Cross Eyes" and "Walleyes" will be retained.

RECOMMENDED ACTION

☐ CONCUR ☐ NO INTEREST

OBJECT (State reasons in full detail, including potential impact upon input or retrieval operations showing significant loss of information. Cite authorities as appropriate.)

Signed

Vocabulary Coordinator _____ Organization _____

RETURN PRIOR TO October 12, 1973

To: ERIC Processing and Reference Facility
ATTN: Lexicographer
4833 Rugby Avenue, Suite 303
Bethesda, Maryland 20014

THESAURUS

ERIC

TERM CHANGE

NOTICE

No. 6

1. PROPOSED CHANGE

Delete the descriptors HORIZONTAL TEXTS and VERTICAL TEXTS.

2. IMPACT

a. POSTINGS BEFORE CHANGE (Dec '72, RIE)

Term	Postings
HORIZONTAL TEXTS	1
VERTICAL TEXTS	1

b. POSTINGS AFTER CHANGE

Term	Postings
------	---

3. REASON FOR CHANGE (Include full justification, citing authorities for definitions, usage, and treatment)

These ancient terms were established as descriptors in 1966. They refer to formats of programed texts. Each has been used only one time, and for the same document. They will be replaced by the more generic descriptor PROGRAMED TEXTS for that one document. They will not be retained as UF's.

RECOMMENDED ACTION

☐ CONCUR ☐ NO INTEREST

OBJECT (State reasons in full detail, including potential impact upon input or retrieval operations showing significant loss of information. Cite authorities as appropriate.)

Signed

Vocabulary Coordinator _____ Organization _____

RETURN PRIOR TO October 12, 1973 To: ERIC Processing and Reference Facility
ATTN: Lexicographer
4833 Rugby Avenue, Suite 303
Bethesda, Maryland 20014

EFF 2/1/8 (2)